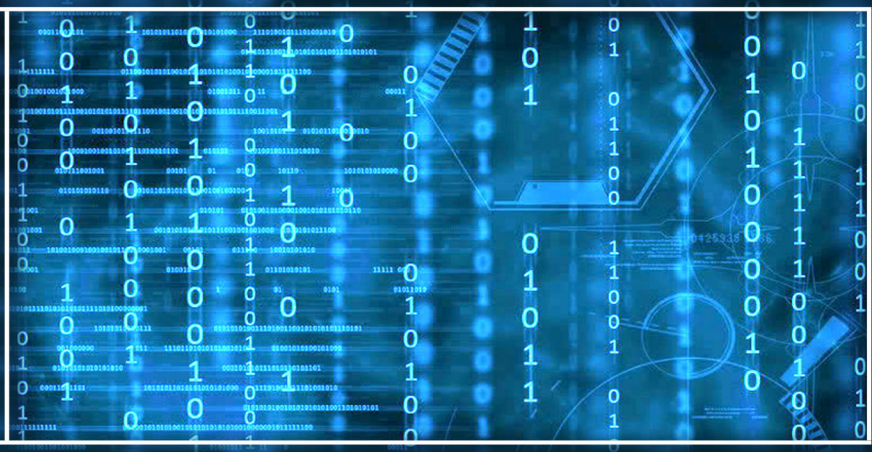
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 13 Issue 2
February 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 2 February 2022
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Random-Valued Impulse Noise Detection and Removal based on Local Statistics of Images

Mickael Aghajarian, John E. McInroy
Department of Electrical and Computer Engineering, College of Engineering and Applied Science
University of Wyoming, Laramie, Wyoming, United States

***Abstract*—Random-valued impulse noise removal from images is a challenging task in the field of image processing and computer vision. In this paper, an effective three-step noise removal method was proposed using local statistics of grayscale images. Unlike most existing denoising algorithms that assume the noise density is known, our method estimated the noise density in the first step. Based on the estimated noise density, a noise detector was implemented to detect corrupted pixels in the second step. Finally, a modified weighted mean filter was utilized to restore the detected noisy pixels while leaving the noise-free pixels unchanged. The noise removal performance of our method was compared with 10 well-known denoising algorithms. Experimental results demonstrated that our proposed method outperformed other denoising algorithms in terms of noise detection and image restoration in the vast majority of the cases.**

***Keywords*—Random-valued impulse noise; noise detection; image restoration; modified weighted mean filter**

I. Introduction

Image noise is an inevitable consequence of some intrinsic (e.g., sensor) and/or extrinsic (e.g., environment) factors such as imperfections in capturing instruments, bit errors in analog-to-digital conversations, malfunctions in camera sensors, and interference in transmission channels. The existence of noise not only degrades the visual quality of images but also adversely affects the performance of image processing and computer vision tasks, like classification, detection, and segmentation. Thus, image denoising is often an essential preprocessing task in the field of image processing and computer vision. The goal of an ideal image denoising method is to remove the noise while maintaining fine structures of images such as edges or corners.

Depending on the sources of noise, image noise can be classified into different categories such as impulse noise, Poison noise, and Gaussian noise. Two common types of impulse noise are the salt-and-pepper (SAP) and random-valued impulse noise (RVIN). In an 8-bit/pixel image, noisy pixels in images corrupted by SAP can take on either the minimum or maximum intensity (i.e., 0 or 255), while for contaminated images by RVIN, corrupted pixels can take any values between 0 and 255. Therefore, detecting noisy pixels contaminated by RVIN is a challenging task. Another challenging issue in detecting the noisy pixels is distinguishing between image edge pixels and corrupted pixels. The big difference between the intensity of image edge and their neighboring pixels might cause noise detectors to falsely detect the image edge pixels as noisy pixels.

Although many algorithms have been proposed for the noise removal problem, there is still room for improvement [1]–[4]. Particularly, for the RVIN removal problem, various methods have been proposed in the literature. The standard median filter (MF) [5], [6] is a widely used nonlinear filter due to its simplicity and high computational efficiency; however, it does not work well for high levels of noise and eliminates fine structures of images, and this leads to blur. In order to improve its performance, some modifications to MF have been proposed. The weighted median filter (WMF) [7] gave more weight to some pixels within the sliding window. It allowed a degree of control of the smoothing by which more image details could be preserved; however, finding suitable weights for different images was not an easy task. Center weighted median filter (CWMF) [8] was a special case of WMF which gave more weights only to the central pixel of the sliding window. The adaptive weighted median filter (AWMF) [8] was another modification to MF in which the filter weights were adapted accordingly based on the local statistics. AWMF could suppress multiplicative noise as well as additive white and impulse noise. Adaptive median filter (AMF) [9] was another method with variable sliding window size.

One characteristic of RVIN is that depending on the noise density, only some parts of image pixels are corrupted while the rest are noise-free. The main drawback of the aforementioned filters is that they restore the entire image by processing all pixels without considering whether the pixel is noisy or not. As a result, they eliminate fine details of images like edges or corners, and this leads to blur. In order to overcome this drawback, several two-step methods have been proposed that are integrated with noise detectors. In the first step, the noise detector determines whether the pixel is corrupted or not. In the second step, only the noisy pixels are restored while other pixels remain unchanged. By doing so, more image details can be preserved and in turn the quality of restored images can be improved. It should be noted that the performance of these methods heavily depends on the proper detection of noisy pixels that is a challenging task for RVIN.

The switching median filter (SMF) [10] calculated the absolute value of difference between the center pixel of the sliding window and the median. If the difference was greater than a predefined threshold, it detected the pixel as noisy and restored it by using the median filter; otherwise, it left the pixel unchanged. The first drawback of SMF is that it uses a fixed threshold to detect noisy pixels. The second drawback is that it restores corrupted pixels using the median value of the current sliding window that might include other noisy pixels, so its

performance for high levels of noise can be deteriorated. Several modifications have been developed to improve the performance of SMF. Noise adaptive soft-switching median filter (NASM) [11] used fuzzy logic to categorize image pixels into four categories named uncorrupted pixel, isolated impulse noise, non-isolated impulse noise and edge pixel. Then, depending on the pixel's characteristic, an appropriate filter (i.e., MF or proposed fuzzy WMF) was utilized to restore corrupted pixels. Adaptive impulse noise detector using CWM (ACWM) [12] utilized the differences between the center pixel of the sliding window and the output of CWM with varied center weights to detect corrupted pixels. In [13], an impulse noise detection technique for SMF was proposed (SWM) that was based on the minimum absolute value of four convolutions obtained using one-dimensional Laplacian operators. SMF with boundary discriminative noise detection for extremely corrupted images [14] used two boundaries by which image pixels were classified into three groups named lower intensity impulse noise, uncorrupted pixels, and higher intensity impulse noise. Then, a modified NASM was utilized to restore corrupted pixels. Directional weighted median filter (DWM) [15] detected the noisy pixels based on the difference between the center pixel and its neighbors aligned with four main directions. Then, WMF was applied iteratively to restore the noisy pixels. In each iteration, the threshold decreased until the maximum number of iterations was reached. In [16], SMF was modified by adding one more noise detector based on the rank order arrangement of pixels in the sliding window. Adaptive switching median filter (ASMF) [17] was another modification to SMF in which the threshold was computed locally from pixels inside the sliding window.

In recent years, some effective noise removal algorithms with local statistics-based impulse noise detectors have been developed. A new statistic based on the Rank-Ordered Absolute Difference (ROAD) [18] was introduced that represented how impulse-like a pixel was in the sense that the larger the impulse, the greater the ROAD value. Then, by incorporating this statistic into a bilateral filtering, a new nonlinear filter was proposed (trilateral filter) which could remove both Gaussian and impulse noise. The Rank-Ordered Logarithmic Difference (ROLD) [19] was developed to improve the performance of the ROAD statistic by identifying more noisy pixels with less false hits. By combining it with an edge-preserving regularization (EPR), ROLD-EPR method was implemented to remove RVIN. A partial differential equation-based image denoising method for random-valued impulse noise (NSDD) [20] was proposed in which two controlling functions were used to distinguish between edge pixels, noisy pixels, and interior pixels. In [21], a detection algorithm for RVIN (ODM) was developed that calculated the standard deviation in different directions in the filtering window. Once the optimal direction was found, a pixel was detected as noise-free if it was similar to pixels in the optimum direction. A fuzzy weighted NLM filter (FWNLM) [22] was implemented that was able to remove RVIN and mixed Gaussian-RVIN. Based on the fuzzy weighting function, the more a pixel was contaminated, the less it was used to restore images. In [23], a new WMF with a two-phase noise detector was proposed. In the first phase, the Rank-Ordered Difference of ROAD (ROD-ROAD) was introduced in which a fuzzy rule was used to detect noisy pixels. In the second phase, another image statistic (minimum edge pixels difference) was proposed to distinguish between edge pixels and noisy candidates. To restore the corrupted images, an iterative denoising algorithm was utilized by combining the proposed two-phase noise detector and the new WMF. A new image denoising method (ℓ_0TV-PADMM) [24] was implemented that was based on the total variation (TV) with ℓ_0-norm data fidelity. Since the resulting optimization problem was non-convex and non-smooth, it was first reformulated as an equivalent mathematical program with equilibrium constraints and then it was solved using a proximal Alternating Direction Method of Multipliers (PADMM). In [25], a new two-phase denoising algorithm (DPC-INR) was implemented using dissimilar pixel counting. In the detection phase, the average difference scheme was used to distinguish whether two neighboring pixels were similar or not, and then the number of dissimilar pixels was compared with a threshold to determine whether the current pixel was noisy. In the filtering phase, an extended trilateral filter was utilized to restore noisy images. An adaptive rank-ordered impulse detector based on local statistics (AROPD-EPR) [26] was introduced in which a piecewise power function was applied to the rank-ordered statistic to enlarge the difference between noisy pixels and noise-free pixels. By combining the noise detector with an improved EPR filter, an effective two-stage iterative denoising algorithm was implemented to remove RVIN.

In this paper, an efficient RVIN removal method was proposed that consisted of three steps (i.e., noise density estimation, noise detection, and image restoration). We made two main contributions one of which was the noise density estimation. As opposed to most existing denoising methods that assume the noise density is known, our method estimated the noise density with high accuracy. The second contribution was proposing an effective RVIN detector using local statistics. Based on the estimated noise density, a noise detector was implemented to detect corrupted pixels. Finally, a modified weighted mean filter was utilized to restore the detected noisy pixels while leaving the noise free pixels unchanged.

The rest of the paper is organized as follows: Section II briefly reviews the ROAD statistic for detecting RVIN. The proposed method is described in Section III. Section IV presents the experimental results and draws a comparison with other state-of-the-art image denoising methods. Finally, Section V provides the conclusion.

II. Review on the Rank-Ordered Absolute Difference Statistic

The RVIN model with noise probability p can be described as follows where u_{ij} and o_{ij} denote the pixel intensity at location (i, j) of the noisy image and original image, respectively, and n_{ij} denotes the value of the noisy pixel at location (i, j).

$$u_{ij} = \begin{cases} o_{ij}, & with\ probability\ 1-p \\ n_{ij}, & with\ probability\ p \end{cases} \quad (1)$$

Unlike SAP noise that takes on either the minimum or maximum intensity, RVIN can take any values between the

minimum and maximum intensity with equal probability. Thus, detecting noisy pixels corrupted by RVIN is much more difficult than SAP.

The ROAD statistic [18] is a widely used local image statistic for detecting RVIN. Let

$$\Omega_{i,j}(N) = \{(i+s,\ j+t) | -N \leq\ s,t\ \leq N\} \quad (2)$$

denote the set of image coordinates in a $(2N+1) \times (2N+1)$ window centered at (i,j) for some positive integer N. Let $\Omega_{i,j}^{0} = \Omega_{i,j} \backslash (i,j)$ be the same as $\Omega_{i,j}(N)$ without its center coordinate. For each coordinate in $\Omega_{i,j}^{0}$, define $\mathrm{d}_{s,t}(u_{i,j})$ as the absolute difference in intensity of the pixels $u_{i+s,j+t}$ and $u_{i,j}$, i.e.,

$$\mathrm{d}_{s,t}(u_{i,j}) = |u_{i+s,j+t} - u_{i,j}|, \quad for\ -N \leq\ s,t\ \leq \mathrm{N}. \quad (3)$$

After sorting the values of $\mathrm{d}_{s,t}(u_{i,j})$ in ascending order, the ROAD statistic can be defined as follows:

$$ROAD_m(u_{i,j}) = \sum_{k=1}^{m} r_k(u_{i,j}) \quad (4)$$

where $2 \leq m \leq (2N+1)^2 - 2$ and $r_k(u_{i,j})$ is the k^{th} smallest value of $\mathrm{d}_{s,t}(u_{i,j})$.

There should be a big difference between the intensity of most corrupted pixels by RVIN and their neighbors while the intensity of most uncorrupted pixels (even edge pixels) should be close to at least half of their neighboring pixels. Therefore, the ROAD value of noisy pixels should be larger than that of noise-free pixels. As a result, the Road value can be used to detect corrupted pixels by RVIN. For a 3×3 window (i.e., $N = 1$), it is suggested to set the value of m to 4 while for a 5×5 window (i.e., $N = 2$) it is recommended to set the value of m to 12 [18].

III. Proposed Method

The proposed method consists of three steps. The first step is the noise density estimation in which RVIN density is estimated with high accuracy. In the second step, a noise detector is utilized to detect corrupted pixels based on the estimated noise density. In other words, the parameters of the noise detector are determined specifically for each noise level. Finally, in the last step, a modified weighted mean filter (MWMF) is used to restore the detected noisy pixels. Each step is explained in detail in the following sections. It is worth mentioning that there are some parameters in each step whose values are determined by using an evaluation dataset that contains 20 images. The evaluation dataset is a subset of BSDS68 dataset [27]. In order to compare the performance of the proposed method with other denoising algorithms, a different dataset, containing 49 images [28], is used.

A. Noise Density Estimator

In the proposed method, a 3×3 sliding window with $m = 4$ is used to calculate the ROAD value of image pixels. If the ROAD value of a pixel is larger than a pre-defined threshold, the pixel will be considered as noisy; otherwise, the pixel will be considered as noise-free. The evaluation dataset is used to determine the value of the threshold to distinguish between noisy and noise-free pixels. The value of the threshold ($T = 78$) is found by trial and error. By sliding the window over the entire image and counting the number of corrupted pixels, the estimated noise density can be easily computed. For instance, if the number of detected noisy pixels in a 512×512 grayscale image is 52,430, the estimated noise density is about 20%.

B. RVIN Detection

In order to decide whether an image pixel $u_{i,j}$ is noisy, all 24 neighboring pixels within a 5×5 window, centered at $u_{i,j}$, are considered in two stages. In the first stage, we decide whether the center pixel $u_{i,j}$ is a noise candidate or noise-free pixel. All detected noise candidates are considered in the second stage to make sure that they are not edge pixels, falsely detected as noise.

The underlying logic of distinguishing between noisy and edge pixels is that in a 5×5 window that does not contain edge pixels, clean pixels should have close intensities while the intensities of contaminated pixels by RVIN vary considerably. In other words, in a smooth area of an image that does not contain edge pixels, there is only one group of clean pixels whose intensities are close together; however, if 5×5 window contains some edge pixels, there could be two groups of pixels whose intensities might differ greatly and yet both groups could be clean pixels.

In the first stage of the proposed method, we first sort the values of all 24 neighboring pixels in ascending order, and then compute the difference between each two successive elements to generate a vector called diff_sort_vec. We then find two biggest groups in diff_sort_vec whose elements are smaller than a threshold (T_1). We consider the first biggest group as clean pixels, called first_clean_pxls, because all the corresponding pixels to this group have close intensities. If there are at least six pixels in the second biggest group, we consider it as the second clean pixels group, called second_clean_pxls, which means the 5×5 sliding window might contain some edge pixels. Now, if the center pixel, $u_{i,j}$, satisfies any of four following conditions, we detect it as a noise candidate in the first stage. The first condition is that the center pixel, $u_{i,j}$, is not within the range of the first clean pixels group (i.e., first_clean_pxls). The second condition is that the ROAD value of the center pixel ($N = 2$ and $m = 10$) be larger than or equal to a threshold (T_2). For the third condition, let n_{5by5} denotes the number of neighboring pixels, within the 5×5 window, whose absolute difference from the center pixel is smaller than a threshold (T_3). The third condition is met if n_{5by5} is smaller than another threshold (T_3'). The only difference between the third and fourth conditions is that we consider the neighboring pixels within a 3×3 window (n_{3by3}), instead of a 5×5 window, and use different thresholds (T_4 and T_4') in the fourth condition.

In the second stage, if detected noise candidates satisfy all of four following conditions, their status will be changed to noise-free pixels. In other words, the detected noise candidates in the first stage could be edge pixels (i.e., noise-free pixels). The first condition is that the center pixel, $u_{i,j}$, is within the range of the second clean pixels group (i.e., second_ clean_pxls). The second condition is that the ROAD value of the center pixel be smaller than a threshold (T_5). To meet the

third and fourth conditions, n_{5by5} and n_{3by3} must be greater than two thresholds (T_6 and T_7). The RVIN detection algorithm can be summarized as follows:

Proposed RVIN detection algorithm:

Input: 5×5 window (w)

Output: $d \in \{0,1\}$

First stage:

1. Remove the center pixel ($w(3,3)$) of the 5×5 window.

2. Vectorize and sort the remaining 24 pixels in ascending order to generate a vector that we call *sorted_w*.

3. Calculate the difference between each two successive elements in *sorted_w* to generate a vector called *diff_sort_vec*

4. Find two biggest groups of elements in *diff_sort_vec* whose elements are smaller than a threshold (T_1). The first and second biggest groups are called *first_clean_pxls* and *second_clean_pxls*, respectively.

5. Calculate the ROAD value of the center pixel with $N = 2$ *and* $m = 10$

6. Find the number of neighboring pixels (n_{5by5}), within the 5×5 window, whose absolute difference from the center pixel is smaller than a threshold (T_3).

7. Find the number of neighboring pixels (n_{3by3}), within the 3×3 window, whose absolute difference from the center pixel is smaller than a threshold (T_4).

8. If [$w(3,3) \notin first\_clean\_pxls$] OR [ROAD $\geq T_2$] OR [$n_{5by5} \leq T_3'$] OR [$n_{3by3} \leq T_4'$], then $d = 1$, i.e., the center pixel is a noise candidate.

Second stage:

9. If [$d = 1$] AND [length(*second_clean_pxls*) ≥ 6] AND [$w(3,3) \in$ *second_clean_pxls*] AND [ROAD $\leq T_6$] AND [$n_{5by5} \geq T_6'$] AND [$n_{3by3} \geq T_7'$], then $d = 0$, i.e., the center pixel is a noise-free pixel.

C. Image Restoration Method

In this paper, we slightly adapted the modified mean filter (MMF) proposed in [29] to restore contaminated pixels by RVIN. If the estimated noise density is less than 35%, contaminated images were restored once; otherwise, they were restored twice to improve their visual quality. In the original MMF method, the center value of a 3×3 sliding window is replaced by the mean of its four horizontal and vertical neighbors aligned with the four main directions, if only each of which is a noise-free pixel. If all neighbors of a center pixel are noisy, it is necessary to move toward the defined directions shown in Fig. 1 to reach closest four noise-free pixels that we call a, b, c, and d. If a pixel is noisy, it will be replaced with another noise-free pixel according to the flowchart that is shown in Fig. 2. Thereafter, the value of center pixel can be simply calculated as the mean of these noise-free pixels, i.e.

$$P0 = \frac{a+b+c+d}{4} \quad (5)$$

We make two modifications to the original MMF. The first modification is that instead of taking the mean of four noise-free pixels, we take the weighted average of four pixels according to the weights proposed in Fig. 3. In the original MMF, the neighboring pixels can be picked more than once. The second modification is that each neighboring pixel cannot be picked more than once.

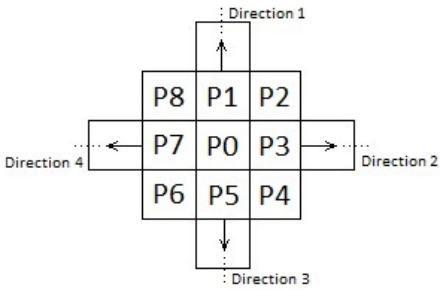


Fig. 1. Four Directions for Selecting Noise-Free Pixels [29].

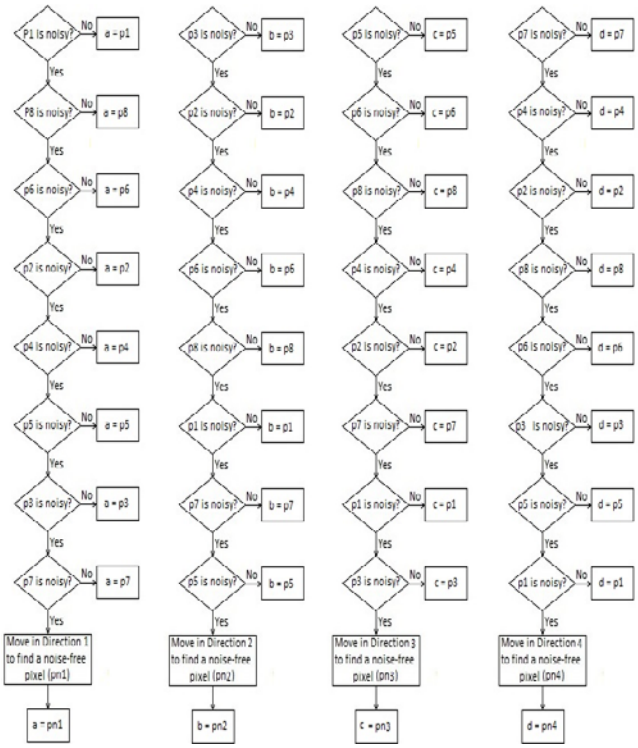


Fig. 2. Flow Chart for Selecting Noise-free Pixels [29].

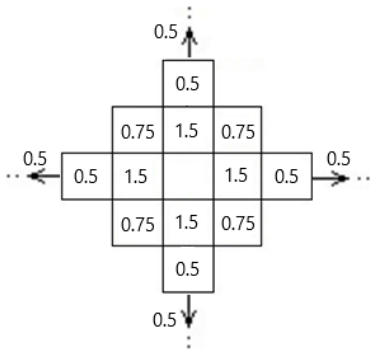


Fig. 3. Assigned Weights to Neighboring Pixels.

IV. Results and Discussion

To compare the performance of our RVIN removal algorithm with other methods, an image dataset [28] consisting of 49 8-bit/pixel grayscale images of size 512×512 are used. Our method is compared with 10 well-known RVIN removal methods all of which are discussed in the introduction. These

methods are ACWM [12], SWM [13], DWM [15], ROLD-EPR [19], NSDD [20], ODM [21], FWNLM [22], ℓ0TV-PADMM [24], DPC-INR [25], and AROPD-EPR [26].

Three well-known metrics (i.e., recall, precision, and F1-score) are used to evaluate the performance of the detection algorithm. Recall (R) is the ratio of correctly detected noisy pixels to the actual noisy pixels while precision (P) is the ratio of correctly detected noisy pixels to all pixels detected as noisy. F1-score is the harmonic mean of precision and recall that is widely used in the field of information retrieval. Note that higher values of F1-score are better that requires both recall and precision be high. The best value for F1-score is one that can be reached only when both recall and precision are equal to one. These metrics can be calculated using the following formulas.

$$R = \frac{TN}{TN + FF} \tag{6}$$

$$P = \frac{TN}{TN + FN} \tag{7}$$

$$F1 - score = \frac{2}{\frac{1}{R} + \frac{1}{P}} \tag{8}$$

where TN is the number of pixels that are correctly detected as noisy, FF is the number of noisy pixels that are falsely detected as noise-free, and FN is the number of noise-free pixels that are falsely detected as noisy.

Restoration results are quantitatively measured by the peak signal-to-noise ratio (PSNR), mean absolute error (MAE), and two-dimensional correlation coefficient (COR) calculated using the following equations.

$$PSNR = 10\ \log_{10}\left(\frac{255^2}{MSE}\right) \tag{9}$$

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(U_n(i,j) - U_d(i,j)\right)^2 \tag{10}$$

$$MAE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|U_n(i,j) - U_d(i,j)| \tag{11}$$

$$COR = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(U_n(i,j) - \overline{U_n})(U_d(i,j) - \overline{U_d})}{\sqrt{\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(U_n(i,j) - \overline{U_n})^2\right)\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(U_d(i,j) - \overline{U_d})^2\right)}} \tag{12}$$

where U_n and U_d are the noisy and de-noised images of size $M \times N$, respectively.

A. Noise Estimation Results

Table I shows the average noise density estimation over 20 images of the evaluation dataset for each noise level. Evaluation dataset is a subset of BSDS68 dataset [27] that is used to determine the values of the parameters (e.g., thresholds). Table II demonstrates the average noise density estimation over 49 images of the test dataset [28]. As can be seen, the proposed method is able to estimate the noise density with high accuracy.

TABLE I. THE AVERAGE NOISE DENSITY ESTIMATION OVER 20 IMAGES OF THE EVALUATION DATASET

Actual noise density	20%	30%	40%	50%	60%
Estimated noise density	19.3%	28.3%	38%	48.6%	59.7%

TABLE II. THE AVERAGE NOISE DENSITY ESTIMATION OVER 49 IMAGES OF THE TEST DATASET

Actual noise density	20%	30%	40%	50%	60%
Estimated noise density	21.3%	30.2%	39.7%	50%	60.7%

B. Noise Detection Results

The performance of RVIN removal methods with noise detector heavily depends on the proper detection of noisy pixels which is not an easy task for RVIN. We compare the performance of our proposed noise detection algorithm with ACWM, SWM, ROAD, ROLD, ECROAD, and AROPD. The average recall, precision, and F1-score for the images of the test dataset for five noise levels are shown in Fig. 4(a), 4(b), and 4(c), respectively. The results of other methods for noise detection, restoration, and run time comparison are taken from [26] in which the parameters are selected as they are suggested in the original papers.

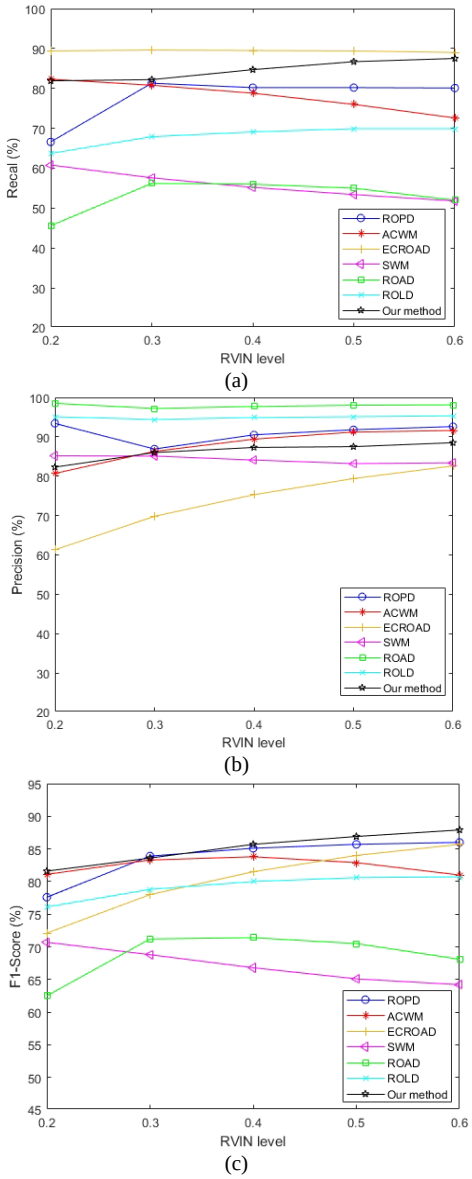


Fig. 4. Comparison of different RVIN Detectors: (a) Average Recall, (b) Average Precision, (c) Average F1-score.

As can be seen from Fig. 4(a), the average recall of our method is lower than that of ECROAD method, but ECROAD method achieved the lowest average precision (Fig. 4b) which means it falsely detects a large number of noise-free pixels as noisy. Similarly, ROAD method achieves the highest average precision and very low average recall which means it wrongly detects a large number of noisy pixels as noise-free. Thus, it is important to consider the F1-score to seek a balance between the recall and precision. As can be seen from Fig. 4(c), except for one noise level (30%), our method achieves the highest average F1-score for all other noise levels indicating the superiority of our noise detection method over other algorithms.

C. Restoration Resultsd

Three well-known metrics (i.e., PSNR, COR, and MAE) are used to quantitatively compare the performance of image restoration. Note that higher values of PSNR and COR are better, whereas lower values of MAE are better. Table III demonstrates the average PSNR of different methods on 49 test images corrupted by five noise densities varying from 20% to 60% with increments of 10. The best average PSNR for each noise density is highlighted in bold. For four noise densities, our method achieved the highest average PSNR while for other noise density (i.e., 60%), it achieved the second best average PSNR. The average COR and MAE are shown in Tables IV and V, respectively. In terms of average COR, our method obtained the best average COR for three noise densities (20%, 30%, and 40%) while for the rest of noise densities, its result is comparable with other methods that achieved higher values for average COR. In terms of average MAE, our method outperformed all other methods for two noise densities (20% and 30%) while for other noise densities, it achieved the second best results. Fig. 5 and 6 demonstrate visual comparison between noise removal methods for the "butterfly" and "bridge" test images corrupted by 55% and 20% RVIN, respectively. It can be seen that our method can remove noise from corrupted images and preserve sharp edges and fine details of the images that yield visually pleasant restoration results.

TABLE III. IMAGE RESTORATION COMPARISON FOR DIFFERENT NOISE DENSITIES OVER 49 TEST IMAGES IN TERMS OF AVERAGE PSNR (DB)

	Noise density				
Method	**20%**	**30%**	**40%**	**50%**	**60%**
ACWM	27.69	26.10	24.36	22.18	19.67
SWM	24.13	21.95	20.08	18.32	16.61
DWM	27.38	26.55	25.60	24.35	22.43
ROLD-EPR	28.23	26.34	25.47	24.75	**23.91**
NSDD	26.82	26.13	25.24	23.96	22.25
ODM	24.84	24.25	23.61	22.95	22.22
FWNLM	27.60	26.63	25.73	24.83	23.62
ℓ0TV-PADMM	22.92	22.16	21.44	20.67	19.89
DPC-INR	26.35	25.40	24.62	23.22	21.00
AROPD-EPR	28.32	26.87	25.73	24.92	**23.91**
Our Method	**28.44**	**26.92**	**26.08**	**24.97**	23.82

TABLE IV. IMAGE RESTORATION COMPARISON FOR DIFFERENT NOISE DENSITIES OVER 49 TEST IMAGES IN TERMS OF AVERAGE COR

	Noise density				
Method	**20%**	**30%**	**40%**	**50%**	**60%**
ACWM	0.9567	0.9469	0.9290	0.8920	0.8183
SWM	0.9263	0.8901	0.8395	0.7678	0.6689
DWM	0.9526	0.9486	0.9405	0.9257	0.8936
ROLD-EPR	0.9602	0.9441	0.9352	0.9269	0.9151
NSDD	0.9542	0.9476	0.9394	0.9233	0.8965
ODM	0.9273	0.9209	0.9115	0.8998	0.8845
FWNLM	0.9563	0.9493	0.9408	**0.9305**	0.9128
ℓ0TV-PADMM	0.9438	0.9305	0.9156	0.8966	0.8727
DPC-INR	0.9466	0.9363	0.9282	0.9079	0.8560
AROPD-EPR	0.9600	0.9493	0.9397	0.9303	**0.9153**
Our Method	**0.9606**	**0.9498**	**0.9410**	0.9287	0.9118

TABLE V. IMAGE RESTORATION COMPARISON FOR DIFFERENT NOISE DENSITIES OVER 49 TEST IMAGES IN TERMS OF AVERAGE MAE

	Noise density				
Method	**20%**	**30%**	**40%**	**50%**	**60%**
ACWM	4.4238	5.5550	7.1848	9.9144	14.722
SWM	5.8205	8.4421	11.801	16.349	22.491
DWM	4.9712	5.7273	6.7776	8.3646	11.153
ROLD-EPR	4.2650	6.1811	7.3081	8.3750	9.6881
NSDD	7.1198	8.1127	9.0755	10.745	13.562
ODM	6.2370	7.1332	8.2456	9.5681	11.208
FWNLM	5.0753	6.0803	7.2792	8.5974	10.172
ℓ0TV-PADMM	12.394	13.362	14.439	15.689	17.156
DPC-INR	4.5945	6.6458	7.5319	9.1382	12.530
AROPD-EPR	4.0431	5.3842	**6.4563**	**7.6499**	**9.2648**
Our Method	**3.9906**	**5.2600**	6.5046	7.9772	9.6501

D. Run Time

The average run time of the denoising methods on 49 test images is shown in Table VI. All experiments were performed on computers equipped with 3.40 GHz CPU. Although the run time of our method is longer than some other methods, it should be noted that it achieved better noise detection and image restoration results in the vast majority of the cases. The run time for heavy noise corruption (i.e., 40%, 50%, and 60%) is about twice longer than that of the low levels of noise corruption because in our method, highly contaminated images were restored twice to improve the quality of restored images. It is worth noting that the noise detection stage of our method took the significant portion of the run time, so future work would be to implement the noise detection algorithm in an optimized and faster way to reduce the run time. Another way to reduce the run time would be running our algorithm in parallel (i.e., parallel computing).

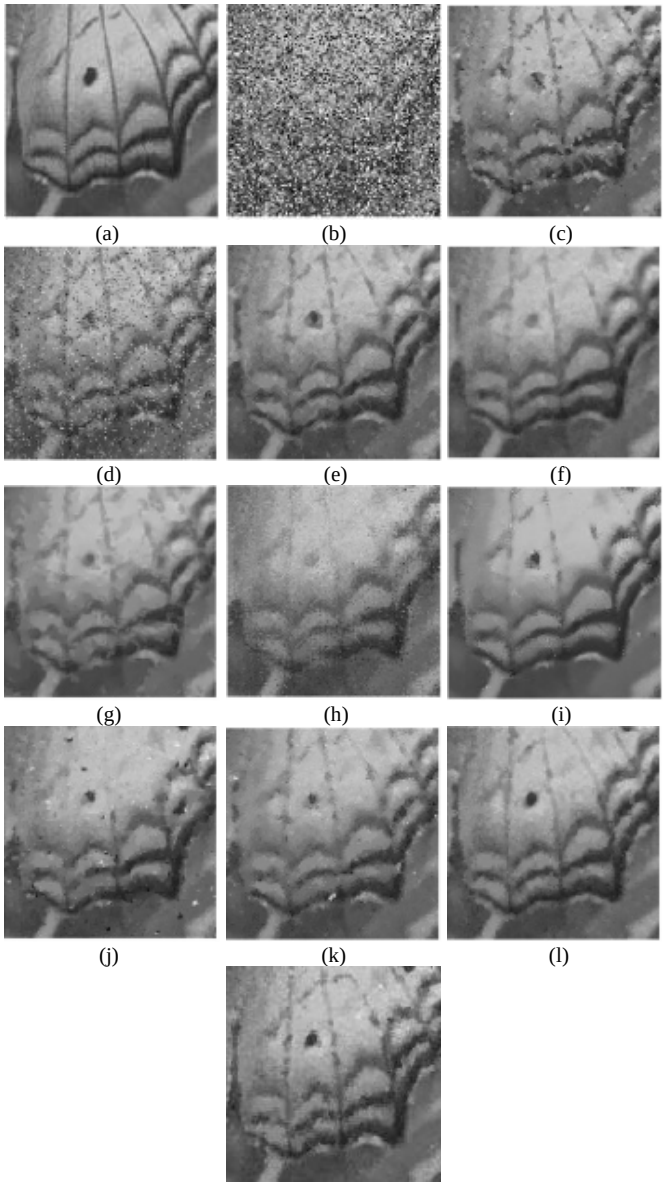
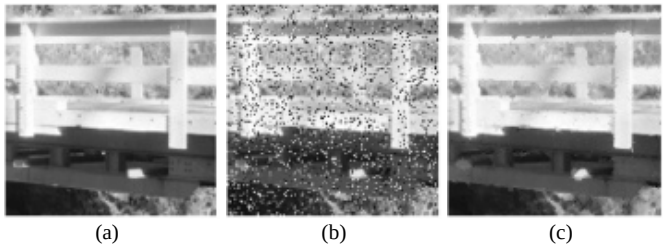


Fig. 5. Comparison of Image Restoration Results of different Methods for Image "Butterfly" Corrupted by 55% RVIN: (a) Clean Image, (b) Noisy Image (55% RVIN), (c) ACWM, (d) SWM, (e) DWM, (f) ROLD-EPR, (g) NSDD, (h) ODM, (i) FWNLM, (j) ℓ0TV-PADMM, (k) DPC-INR, (l) AROPD-EPR, (m) our Method.



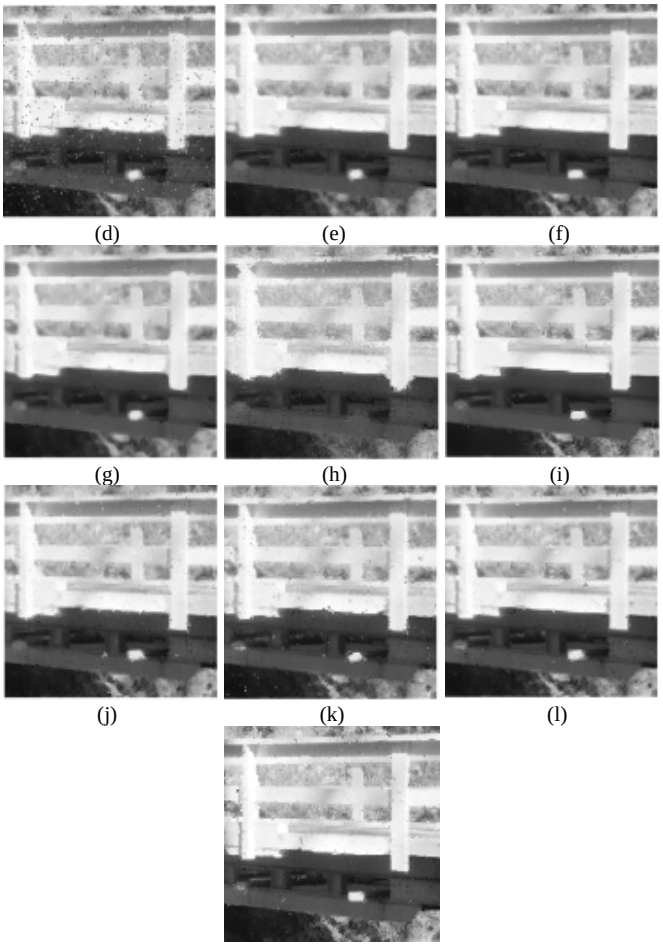


Fig. 6. Comparison of Image Restoration Results of different Methods for Image "Bridge" Corrupted by 20% RVIN: (a) Clean Image, (b) Noisy Image (20% RVIN), (c) ACWM, (d) SWM, (e) DWM, (f) ROLD-EPR, (g) NSDD, (h) ODM, (i) FWNLM, (j) ℓ0TV-PADMM, (k) DPC-INR, (l) AROPD-EPR, (m) our Method.

TABLE VI. COMPARISON OF RUN TIME IN SECONDS

	Noise density				
Method	**20%**	**30%**	**40%**	**50%**	**60%**
ACWM	5.53	5.66	5.60	5.60	5.60
SWM	4.56	4.56	4.48	4.69	4.57
DWM	46.98	45.55	44.82	44.83	46.67
ROLD-EPR	1.63	4.16	5.07	5.52	7.75
NSDD	0.62	0.63	1.31	3.76	5.71
ODM	15.58	14.35	14.46	14.88	15.50
FWNLM	276.39	268.56	277.34	274.38	271.71
ℓ0TV-PADMM	1.38	1.39	1.43	1.49	1.55
DPC-INR	9.15	13.54	14.32	14.71	14.78
AROPD-EPR	1.76	2.82	4.74	7.12	11.25
Our Method	7.16	7.33	13.89	14.19	14.27

V. Conclusion

In this paper, we presented an efficient three-step noise removal method for grayscale images corrupted by RVIN. In the first step, we estimated the noise density of corrupted images with high accuracy. Based on the estimated noise density, a noise detector found corrupted pixels that were restored by using a modified weighted mean filter. In order to evaluate the performance of the proposed method, we drew a comparison with 10 denoising algorithms. In the vast majority of the cases, our method outperformed other algorithms which indicated the effectiveness of the proposed method. Future work would be to implement the noise detection algorithm in a faster way to reduce the run time.

References

[1] B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma, “Image denoising review: From classical to state-of-the-art approaches,” Inf. Fusion, vol. 55, pp. 220–244, 2020.

[2] L. Fan, F. Zhang, H. Fan, and C. Zhang, “Brief review of image denoising techniques,” Vis. Comput. Ind. Biomed. Art, vol. 7, 2019.

[3] K. Zhang, W. Zuo, and L. Zhang, “FFDNet: Toward a fast and flexible solution for CNN based image denoising,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4608–4622, 2017.

[4] M. Aghajarian, J. E. McInroy, and S. Muknahallipatna, “Deep learning algorithm for Gaussian noise removal from images,” J. Electron. Imaging, vol. 29, no. 04, p. 1, 2020.

[5] W. K. Pratt, “Median filtering,” Los Angeles, 1975.

[6] R. E. Gonzalez, Rafael C., Woods, Digital Image Processing. 2002.

[7] D. R. K. Brownrigg, “The Weighted Median Filter,” Commun. ACM, vol. 27, no. 8, pp. 807–818, 1984.

[8] S. J. Ko and Y. H. Lee, “Center Weighted Median Filters and Their Applications to Image Enhancement,” IEEE Transactions on Circuits and Systems, vol. 38, no. 9. pp. 984–993, 1991.

[9] H. Hwang and R. A. Haddad, “Adaptive Median Filters: New Algorithms and Results,” IEEE Trans. IMAGE Process., vol. 4, no. 4, pp. 499–502, 1995.

[10] T. Sun and Y. Neuvo, “Detail-preserving median based filters in image processing,” Pattern Recognit. Lett., vol. 15, no. 4, pp. 341–347, 1994.

[11] H. L. Eng and K. K. Ma, “Noise adaptive soft-switching median filter,” IEEE Trans. Image Process., vol. 10, no. 2, pp. 242–251, 2001.

[12] T. Chen and H. R. Wu, “Adaptive impulse detection using center-weighted median filters,” IEEE Signal Process. Lett., vol. 8, no. 1, pp. 1–3, 2001.

[13] S. Zhang and M. A. Karim, “A New Impulse Detector for Switching Median Filters,” IEEE Signal Process. Lett., vol. 9, no. 11, pp. 360–363, Nov. 2002.

[14] P. Ng and K.-K. Ma, “A Switching Median Filter With Boundary Discriminative Noise Detection for Extremely Corrupted Images,” IEEE Trans. Image Process., vol. 15, no. 6, pp. 1506–1516, 2006.

[15] Y. Dong and S. Xu, “A new directional weighted median filter for removal of random-valued impulse noise,” IEEE Signal Process. Lett., vol. 14, no. 3, pp. 193–196, 2007.

[16] C. C. Kang and W. J. Wang, “Modified switching median filter with one more noise detector for impulse noise removal,” AEU - Int. J. Electron. Commun., vol. 63, no. 11, pp. 998–1004, 2009.

[17] S. Akkoul, R. Lédée, R. Leconge, and R. Harba, “A new adaptive switching median filter,” IEEE Signal Process. Lett., vol. 17, no. 6, pp. 587–590, 2010.

[18] R. Garnett, T. Huegerich, C. Chui, and W. He, “A universal noise removal algorithm with an impulse detector,” IEEE Trans. IMAGE Process., vol. 14, no. 11, pp. 1747–1754, 2005.

[19] Y. Dong, R. H. Chan, and S. Xu, “A detection statistic for random-valued impulse noise,” IEEE Trans. Image Process., vol. 16, no. 4, pp. 1112–1120, 2007.

[20] J. Wu and C. Tang, “PDE-Based Random-Valued Impulse Noise Removal Based on New Class of Controlling Functions,” IEEE Trans. Image Process., vol. 20, no. 9, pp. 2428–2438, 2011.

[21] A. S. Awad, “Standard deviation for obtaining the optimal direction in the removal of impulse noise,” IEEE Signal Process. Lett., vol. 18, no. 7, pp. 407–410, 2011.

[22] J. Wu and C. Tang, “Random-valued impulse noise removal using fuzzy weighted non-local means,” Signal, Image Video Process., vol. 8, no. 2, pp. 349–355, 2014.

[23] L. Liu, C. L. P. Chen, Y. Zhou, and X. You, “A new weighted mean filter with a two-phase detector for removing impulse noise,” Inf. Sci. (Ny)., vol. 315, pp. 1–16, 2015.

[24] G. Yuan and B. Ghanem, “L0TV-MCP: A New Method for Image Restoration in the Presence of Impulse Noise,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5369–5377.

[25] Z. Shi, Z. Xu, K. Pang, Q. Cao, and T. Luo, “Dissimilar pixel counting based impulse detector for two-phase mixed noise removal,” Multimed. Tools Appl., vol. 77, no. 6, pp. 6933–6953, 2018.

[26] Q. Xu, Y. Li, Y. Guo, S. Wu, and M. Sbert, “Random-valued impulse noise removal using adaptive ranked-ordered impulse detector,” J. Electron. Imaging, vol. 27, no. 01, p. 013001, 2018.

[27] S. Roth and M. J. Black, “Fields of experts,” Int. J. Comput. Vis., vol. 82, no. 2, pp. 205–229, 2009.

[28] J. A. García et al., “COMPOUND GAIN: a visual distinctness metric for coder performance evaluation,”, https://ccia.ugr.es/cvg/CG/base.htm.

[29] M. Aghajarian, J. E. McInroy, and C. H. G. Wright, “Salt-and-pepper noise removal using modified mean filter and total variation minimization,” J. Electron. Imaging, vol. 27, no. 01, p. 013002, 2018.

Robotic Ad-hoc Networks Connectivity Maintenance based on RF Signal Strength Mapping

Mustafa Ayad[1], Mohamed Ayad[3]
Electrical and Computer Engineering
The State University of New York at Oswego
Oswego, NY, USA

Richard Voyles[2]
Electrical and Computer Engineering Technology
Purdue University at West Lafayette
West Lafayette, IN, USA

***Abstract*—Network connectivity preservation is one of the substantial factors in achieving efficient mobile robot teams' maneuverability. We present a connectivity maintenance method for a robot team's communication. The proposed approach augments the Radio Frequency Mapping Recognition (RFMR) method and the signal strength gradient decent approach for an overall goal to create a Proactive Motion Control Algorithm (PMCA). The PMCA algorithm controls and helps strengthen mobile communicating robots' connectivity in the existent Radio Frequency (RF) obstacles. The RFMR method takes advantage of Hidden Markov Models (HMMs) results, which assist in learning electromagnetic environments depending on measurements of RF signal strength. The classification results of HMM lead the robots to resolve whether to continue the current trajectory for avoiding the obstacle shadow or move back to desirable robust Signal Strength (SS) positions. In both cases, the robot will run the gradient approach to determine the signal change trend and drive the robot toward the strong SS direction for maintaining link connectivity. The PMCA, depending on the results of RFMR and gradient approaches, promises to preserve robots' motion control and link connectivity maintenance.**

***Keywords*—RF mapping recognition; link connectivity; gradient algorithm**

I. Introduction

The majority of Communication networks, especially wireless networks, are deployed in territories with different interference sources (Different obstacles), affecting the communication signals and creating no Line Of Sight (LOS) among communication devices, so they can not identify each other. However, the Frezonet zone where the signal propagates should be free of interferences sources such as conducting and conducting obstacles of different types to an actual LOS [1,2]. One problem of the RF communications in disasters such as crumpled buildings is many signal interference sources that cause no LOS and disrupt the communication signal. Robot swarms of small size can collaborate in search and rescue environments and accomplish tasks that no one robot can complete alone [3]. Fig. 1 illustrates the urban search and rescue (USAR) robot team collaborating and communicating to transmit data to the network base station (BS). The robot team will encounter many problems when discovering the collapsed area. One of the critical problems is maintaining a reliable link between the robot team members to transmit the message to the BS. For example, a single robot could not send messages directly from the most distant network topology to the BS. What's more, each robot in the team has different duties. For example, it searches for survivors, maintains communication through network topology, and transmits data to the BS.

Collaborating teams of small robots can facilitate tasks beneficial to monitoring, surveillance, and other rescue services in unsafe locations [4]. However, they have limited mobility, power, and communication coverage [5]. Consequently, the system resources are distributed among multiple robots, which work as a team to accomplish a mission. Hence, each small robot in the collaborating team has inadequate sensing and processing abilities for the assigned tasks, e.g., mapping the collapsed area, transmitting acquired data to BS, and carrying necessary sensors for the mission.

A robot in the robotic network can quickly lose communication with team members while collaborating. Therefore reliable strategies for wireless communication are essential [6]. Consequently, a dedicated link maintenance strategy is vital for reliable mobile ad hoc networks (MANETs) connectivity, particularly when the network experiences sporadic connectivity caused by hostile environments. Hence, network connectivity maintenance is a target for achieving adequate network performance. In this context, it is possible to employ the variations in the SS measurements in control algorithms that control motion and preserve connectivity.

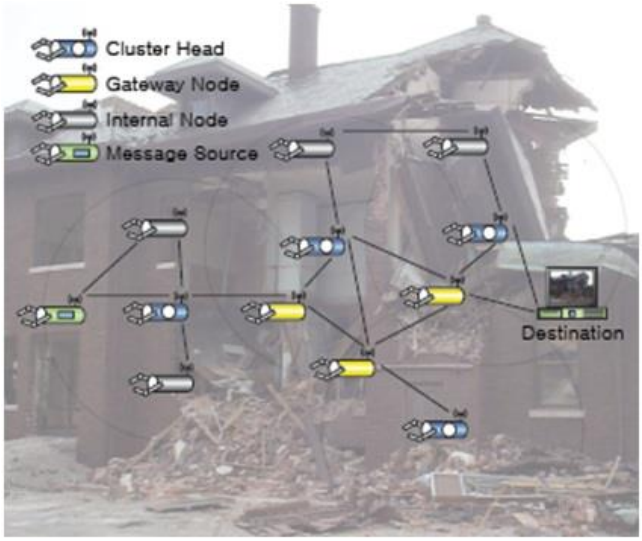


Fig. 1. Robots Warm in a Collapsed Building.

Furthermore, significant developments in robotic networks have led to reliable, self-organizing communication schemes that do not require the collaboration of robots with existing communication infrastructure. Besides, the communication scheme has given bonds to the motion control concept for movable agents, tied to the robot's movement ability to reach proper positions in the field to preserve connectivity and accomplish assigned tasks [7,8]. In [9], the authors manifest the possibility of designing trajectories by co-optimizing sensing and communication information when motion planning.

This article introduces the RFMR method, which uses SS measurements collected from the field to discover, identify, and identify hostile environments with different RF obstacles. In addition, this method also studies the impact of RF obstacles on RF-SS in various scenarios. In addition, according to the RFMR method, we propose a SS gradient algorithm to control the movement trend of the robot. Most up-to-date gradient strategies use a standing interference source to compute a two-dimensional (2-D) gradient to an appointed stable source of the signal source, similar to regression in a 2-D space. However, these approaches did not extend their gradient calculation to nodes in motion that require a four-dimensional (4-D) space gradient estimation. Consequently, augmenting the gradient algorithm and RFMR approach will support creating the PMCA control algorithm to preserve mobile robots' successful communication in the appearance of RF obstacles.

In the simulation and physical experiments, two robots are moved in a different direction around RF obstacles and study their effect on the measurements of the RF signal. The interference sources include cages, walls, and cylinders of various dimensions. The obstacles used are conductively made of a Perfect Electrical Conductor (PEC). When robots move and collect SS measurements around a fixed position obstacle, the collected SS string along the robot's path retains useful information for obstacle recognition and classification. First, the collected SS measurements were segmented, and then features (observation sequence) were extracted using Fast Fourier Transform FFT [4,5]. Afterward, the observation sequences are coded using a clustering algorithm known as K-means [10]. Then, first-order HMMs are used to model the observation sequences [11,12], trained, and then used for the RFMR method. Using this approach, the outcomes of the experiments show very accurate recognition results. As the movable robots identify the nature and assess the dimensions of the confronted obstacle, the PMCA will decide whether to continue moving along the current trajectory to bypass the RF obstacle's shadow or reverse drive to a position where the robot gains a reliable SS. In either case, the gradient descent algorithm is applied, and the multi-dimensional gradient of the strong SS direction used by the robot PMCA for connection maintenance is estimated.

In short, our proposed robot PMCA for preserving communication links and fixing disrupted links is implemented depending on RFMR and gradient methods. The RFMR method uses HMM to discover the RF environment based on SS measurements to estimate the type and size of obstacles. The gradient algorithm outcome decides the direction of the robust SS to maintain connectivity. Finally, the robot motion control can keep the connection and repair the broken link depending on the RFMR and gradient results. The PMCA algorithm's reliability and performance were tested by conducting various simulation experiments. Consequently, the proposed approach has exhibited assuring solutions for the connectivity problem of a robotic network.

We organized this paper as follows. The relevant prior work in a controlled mobile sensor network, estimating and mapping radio signals, is briefly introduced in Section II. Section III presents the RFMR method formulation and modeling to justify this new development. In Section IV, the physical obstacle experiments are described. The simulation and physical results validation is described in Section V. Section VI explains the obstacle parameterization. In Sections VII and VIII, RFMR based on HMM and numerical results are explained. PMCA and gradient methods are described in Sections IX and X. The experimental gradient results are presented in Section XI. The control motion algorithm simulation is illustrated in XII. Section XIII presents conclusions and future work.

II. Literature and Related Work

Recently, connectivity and SS measurements have become essential attributes of communication networks to ensure quality communication [13,14]. In addition, the robot network should maintain connectivity when performing tasks [15]. Based on the information from radio SS, authors in [16] calculate the 2-D gradient of a robot in motion. Besides, the authors calculated the gradient of the robot in mobility to a stationary source of RF signal. In [17], the possibility of localizing and navigating to a standstill source of RF signal by utilizing the two-dimensional gradient of a cooperating sensor network is studied. Authors [15,17] defined a 2-D gradient for a robot in motion to a standstill source of RF signal. The robot follows a predefined trajectory to accommodate its velocity. Authors in [18] proposed a probabilistic framework for evaluating wireless channels. Authors in [19] developed tools for estimating and mapping radio signals. In an attempt to create an urban radio map, Authors in [20] constructed a BS in an unknown location, which transmits data to one or more mobile robots to create a map of the radio signal for a specified area. An algorithm that sets the team's goals and controls its movement makes sure it reaches designated targets without degrading the quality of the link maintaining the map.

Moreover, [20] discussed experimental validation of a procedure that automatically conserves the connection between collaborating robots over such a distributed network. A feedback control framework that is distributed and does not impose restrictions on the network's structure except for desired connectivity specifications has been proposed by [21] concerning the local connectivity of a network. In [22], the authors introduce a measure that provides a measure of the network's global connectivity if certain conditions are met. The authors [22] solved stratum stability's distributed maintenance problem with the nearest neighbor links. Authors propose robots to overcome environmental interference and enable end-to-end communication [23,24]. Several measurements in the robot network are used to estimate the

spatial variation of the wireless channel by [25], where the link quality predicts communication.

Current research on wireless sensor networks focuses on developing energy-saving routing protocols, distributed data compression, transmission schemes, and cooperative signal processing algorithms [26]. In addition, our research is interested in creating a wireless video sensor network of robots that work in hazardous areas and accomplish different tasks while maintaining team connectivity. The wireless network of video sensors is a locally distributed mobile sensor system that captures, processes, and transmits information through a self-organizing wireless network, as shown in Fig. 2. Compared with traditional communication systems, wireless video sensor networks operate underneath a unique set of resource restrictions, including airborne computing and transmission bandwidth. In [12], the authors investigated the resource utilization behavior and analyzed the Video sensor network performance under resource constraints.

III. RFMR Formulation and Modeling

The RFMR Method depends on the RF-SS determinations on the robot's path. First, the technique identifies and classifies the types of RF shadows on the robot's path. Then, it provides the learned knowledge to PMCA. The outcome of the HMM gained from the RFMR method advises the moving robots underneath the obstacle's special effects. After that PMCA relies on HMM results to determine the proper control on the robot motion, firstly, to recover from the shadow of RF obstacles and then preserve the connectivity of the robot. PMCA decided to let the robot move forward on the current trajectory under the influence of the shadow of the RF obstacle. It did that depending on the size and type of the obstacle. In contrast, PMCA guided them to back movement to a vital SS location and then applied the SS gradient algorithm to find the trend of another robot to communicate.

All RFMR experiments use two mobile sensors transmitting and receiving RF signals at a frequency of 2.4 GHz. They measure the RF-SS at their present location. Multipath, fading, and interference may affect the measured RF signal [3,29]. Mobile sensors, at t=0, are positioned at (x_t, y_t) in the 2-D Cartesian space. The mobile sensors 2-D configuration spaces are divided into grids of equal area. The grid width is Δx = Lx / M, where Lx is the length and M is the number of segments, lengths, and segments along the x-axis. The grid length is Δy= Ly / N, where Ly and N are the lengths and the number of segments in the y-axis. For example, the grid's width might be $\frac{1}{3}\lambda, \frac{2}{3}\lambda$ or λ, which is *12.5 cm* at *2.4 GHz.*

In RFMR simulation experiments, the robots move predefined trajectories to acquire RF-SS measurement. The robot's trajectory, *l* th, can be expressed by.

$$x_{t,l}^{(i)} = x_{0,l}^{(i)},\ y_{t,l}^{(i)} = y_{0,l}^{(i)} + k\,\Delta_y\,, t\ 1,2,\dots.,N, \quad (1)$$

Where the trajectory index is *l*, the robots index is $\boldsymbol{i}\ \varepsilon\{1,2\}$, the *i*th robot start location at time *t=0* is $(x_{0,l}^{(i)}, y_{0,l}^{(i)})$. The *i*th robot's motion starts at the initial location $(x_{0,l}^{(i)}, y_{0,l}^{(i)})$ and increased by a step size of $\boldsymbol{\Delta_y}$ along with the *y-direction* is defined in Equation (1). Besides, the first robot started at $(x_{0,l}^{(1)} = l\Delta_x, y_{0,l}^{(1)} = 0)$ for the trajectory, ***l*th**. Then, the second robot location is expressed in $x_{0,l}^{(2)} = x_{0,l}^{(1)} + d$, and $y_{0,l}^{(2)} =, y_{0,l}^{(1)}$. The robots have the exact coordinates in the y-axis, and they are a d distance in the x-***axis***. Fig. 3(a) shows an experiment scenario of two robots. The SS at the receiver antenna can be expressed as.

$$S_i^{(j)}(t) = f\left(x_{0,l}^{(1)}, y_{0,l}^{(1)}, x_{t,l}^{(1)}, y_{t,l}^{(1)}, x_{0,l}^{(2)}, y_{0,l}^{(2)}, x_{t,l}^{(2)}, y_{t,l}^{(2)}\ \phi j\right) \quad (2)$$

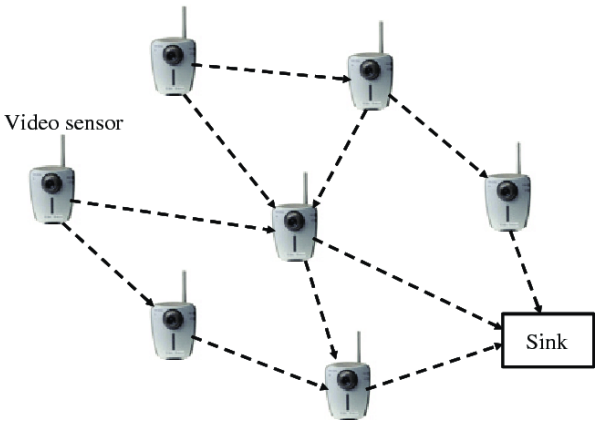
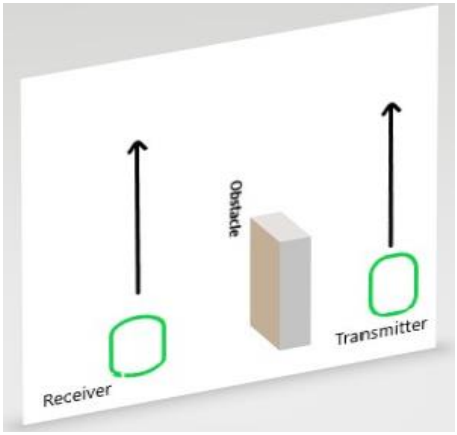
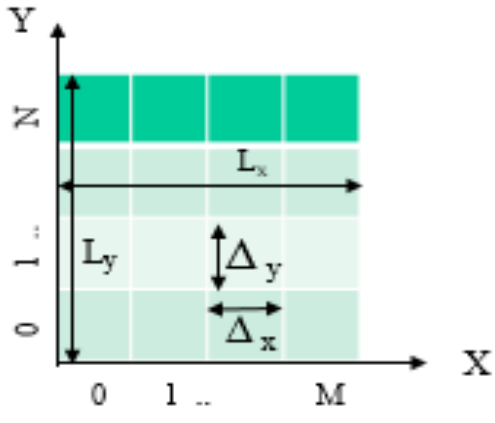


Fig. 2. Wireless Video Sensor Network.



(a)



(b)

Fig. 3. (a) A Transmitter and Receiver Exchange about an Obstacle (b) The 2-D configuration Space.

This is the SS measurement on the trajectory, *l*th, in the RF obstacle of type ***j*** existence at time *t*. The $S_i^{(j)}(t)$ function represents the robot's start location $(x_{0.l}^{(i)}, y_{0.l}^{(i)})$, time t robot's location $(x_t^{(i)}, y_t^{(i)})$ and the special characteristics of the obstacle ϕ_i. The index of trajectory l is $l = 1,\ldots, L^{(j)}$ for each obstacle of type *j*. $L^{(j)}$ is the trajectory's number in the existence of type *j* obstacle. In equation (2), $j \in \{1,2,3\}$ characterizes the type of the obstacle and $\phi j = \{ (x_c^{(j)}, y_c^{(j)}), \Theta^{(j)}\}$ signifies the obstacle's characteristic and comprises the obstacle shape parameters $\Theta^{(j)}$ (dimensions information) and the obstacle central position $(x_c^{(j)}, y_c^{(j)}), e.g.$ Wall obstacle has a central position ($^{Lx}/_2$, $^{Ly}/_2$), [34].

We demonstrated the SS measurements in the field, expending three different types of RF obstacles. Therefore, it can classify and identify the RF characteristics of a particular type of RF obstacle by examining the changes in the SS measurements obtained at diverse locations from different trajectories [27,28]. Computer Simulation Technology (CST) is used for the simulation experiment. It is a professional 3D electromagnetic simulation tool [29]. The simulation uses a 60mm x 60mm patch antenna. It sends and receives communication signals and creates interference from purely conductive materials [30].

A. *Wall Obstacle*

One of the known obstacles of various dimensions (7 x 30 x 30 cm^3, 10 x 30 x 30 cm^3, and 15 x 130 x30 cm^3) are used in the experiments. The RF-SS result in the field is shown in Fig. 4(a) for the 10 x 30 x 30 cm^3 wall. When the transmitter approaches the wall's edge on one side and the receiver is one meter far on the other side, SS drops down and becomes very low, and vice versa. The SS improves as the receiving or transmitting robot moves away from the obstacle edges, as depicted in Fig. 4(a). The top view of the results is illustrated in Fig. 4(b), where the dark red dots illustrate spikes of Fig. 4(a). Fig. 4(c) depicts various waveforms resulting from the wall obstacles at the receiving robot location for different distances. The waveform reflects the influence of RF obstacles on the RF-SS between robots when the robot moves about the RF obstacle.

B. *Cage Obstacle*

A Faraday-like cage is shown in Fig. 5(a). It is made of PEC material. The size of the cage is 30 x 30 x 30 cm^3. The SS drops and signal expire when a robot is trapped in the cage, as presented in Fig. 5(a). Due to conducting material effects, the SS exhibits oscillating patterns as either antenna approaches the cage opening. The signal drops down when there is no LOS between antennas and becomes weak, as presented in Fig. 5(a). Fig 5(b) is the SS intensity image of Fig. 5(a) and illustrates that the SS goes up as a LOS exists. The effect on RF-SS by the cage obstacle is depicted in Fig. 5(a), 5(b), and 5(c). Different waveforms represent measurement sequences of different trajectories around the obstacle are depicted in Fig. 5(c).

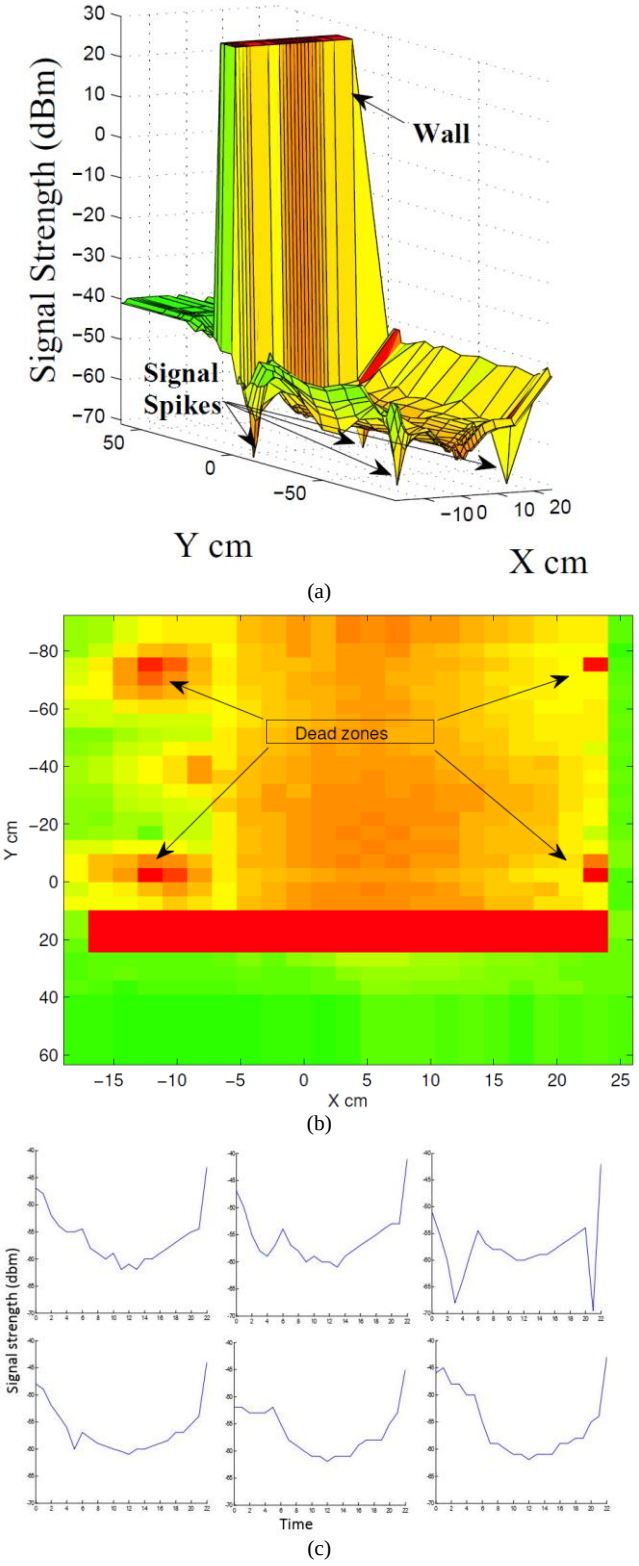


Fig. 4. Wall Obstacle (a) RF-SS Measurements (b) Top view of (a), and (c) Waveforms for Multiple Trajectories.

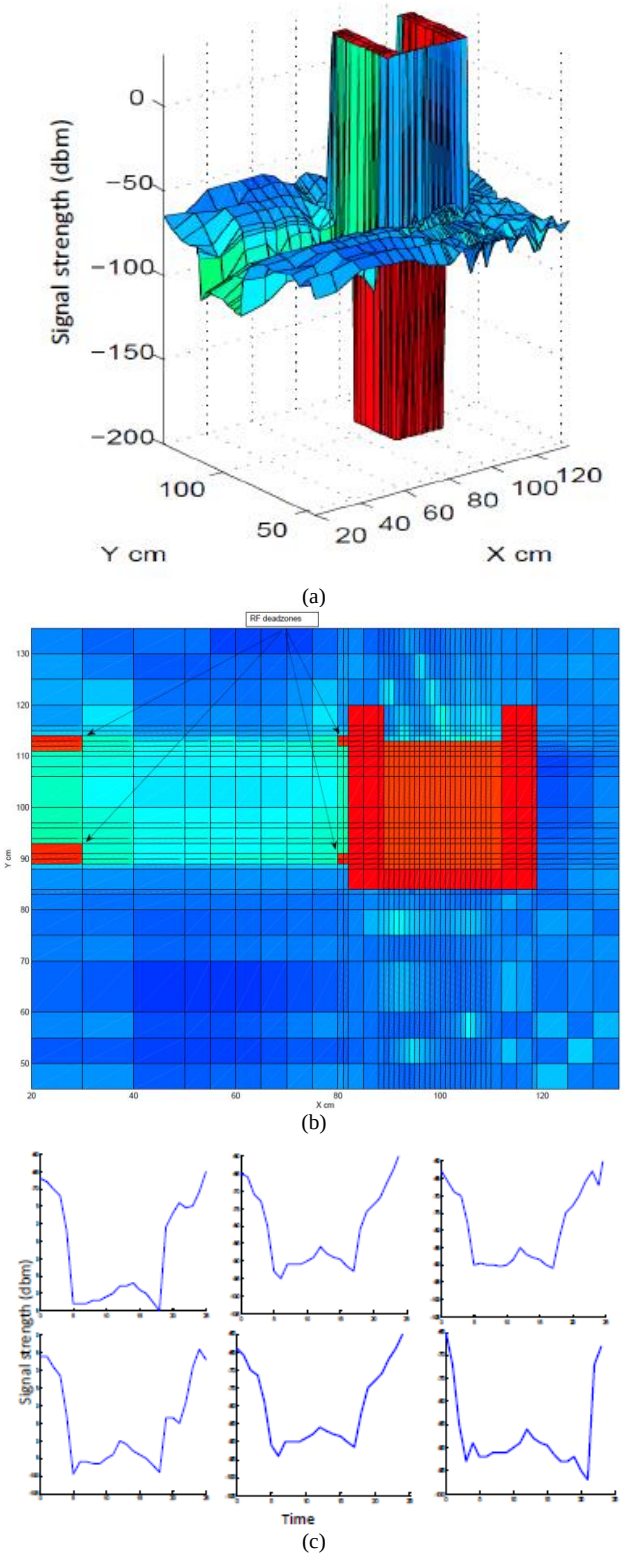


Fig. 5. Cage Obstacle (a) RF-SS Measurements (b) Top view of (a), and (c) Waveforms for Multiple Trajectories.

C. Cylinder Obstacle

Cylinders of radiuses 10 cm, 15 cm, and 20 cm and height of 30cm were used in this experiment.

SS measurements simulation results in the field in the existing on an obstacle of 15 cm diameter centered in the testing area are depicted in Fig. 6(a). The SS dropped down and became unreliable as the receiving robot approached the cylindrical obstacle. Fig. 6(b) presents the SS intensity image of Fig. 6(a).

The influence of the cylinder on the RF-SS is depicted in Fig. 6(a), 6(b), and 6(c). Since the antenna moves in line on either obstacle side, the SS sequence contains enough evidence to identify the obstacle type**.** When the receiver antenna is close to an HF obstacle, the shadow of the HF obstacle in the HFSS measurement will produce different waveforms for different distances.

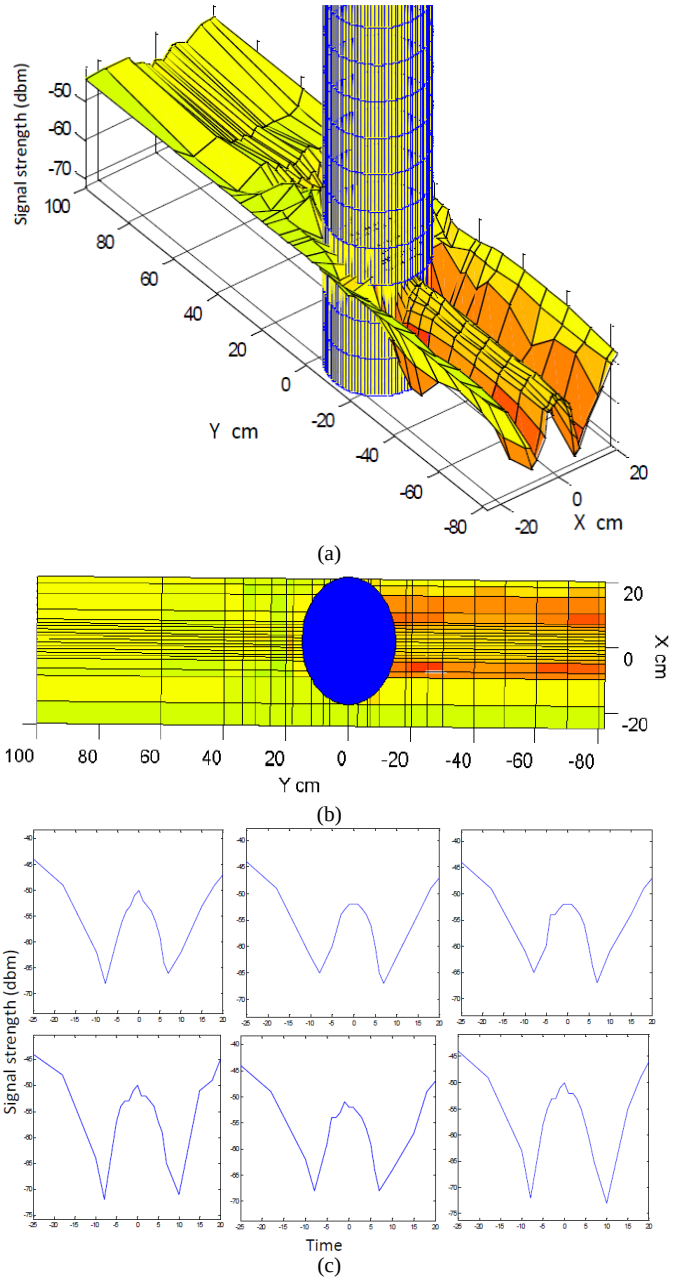


Fig. 6. Cylinder Obstacle (a) RF-SS Measurements, (b) Top view of Fig. 6, and (c) Waveforms for Multiple Trajectories.

IV. Physical Experiments for Multiple Obstacles

Conductive known RF obstacles such as cages, walls, and cylinders are created to run multiple physical experiments in the field [27]. Then, we sought a minimum interference environment to run the experiments, and a CC2510 development kit was used. The copper obstacle is centered on the cardboard box in the laboratory space. Then, 2.4 GHz transceivers are moved manually in all directions around the obstacle. We recorded SS measurements at different antenna positions around the obstacle. RF-SS measurements are made on both sides up to 100 cm in all directions. Next, we made various RF obstacle shapes similar to those used for simulation. The physical results are based on surroundings and floor type. Different materials such as carpets and wood have other effects on the RF signal. Running multiple extensive experiments to choose the best environment leads us to select a box of 15 cm height for best results [27].

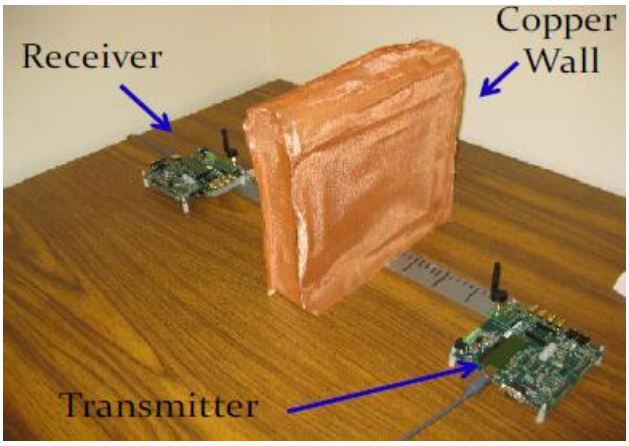


Fig. 7. Copper Wall Obstacle, Receiver and Transmitter.

A. Copper Wall Obstacle

A wooden box of dimensions (10 x 30 x 30 cm^3) is created and then covered with a copper screen, as seen in Fig. 7. We collected SS measurements around the obstacle by moving the antennas in different directions. The results are shown in Fig. 8(a). The SS measurements range from high to low, depending on the obstacle effects. The SS turned out to be deficient, as indicated by the spikes in Fig. 8(a). The SS improved when the transceiver was 1 m apart on one side while the other was still and close to the obstacle, but it remained low as the obstruction prevented LOS.

The physical results presented in Fig. (8) approximates the simulation outcomes shown in Fig. 4(a). The low SS is apparent in Fig. 8(b) that the top view of Fig. 8(a). Therefore, we take advantage of the position of the spikes in estimating the obstacle dimensions, which is valid for simulation results too.

When the transceivers diverge from the obstacle shadow, SS improves, and it reaches the maximum as the transceivers maintain a LOS. The copper obstacle effects on the SS are depicted in Fig. 8(a)-(c). Therefore, after examining the results of the experiment robots' movements all over the obstacle for multiple straight trajectories, it is clear that the SS contains helpful information that is used to recognize and classify obstacle types. Fig. 8(c) illustrates the signature of the obstacle on the RF signals, and it shows different signal shapes for various trajectories.

B. Copper Cage Obstacle

A four-sided wood cage of length = 30 cm, width =30 cm, and height =30 is created and covered by a screen of pure copper. The obstacle is centered on a cardboard box that insulates the transceivers from the ground. Next, the transceivers, which preserve a distance of 1 m apart, are moved around the obstacle. Finally, the transceivers move in all directions outside and inside the cage for SS measurements. Inside the cage, the SS is extremely low and not conducive, as shown in Fig. 9(a).

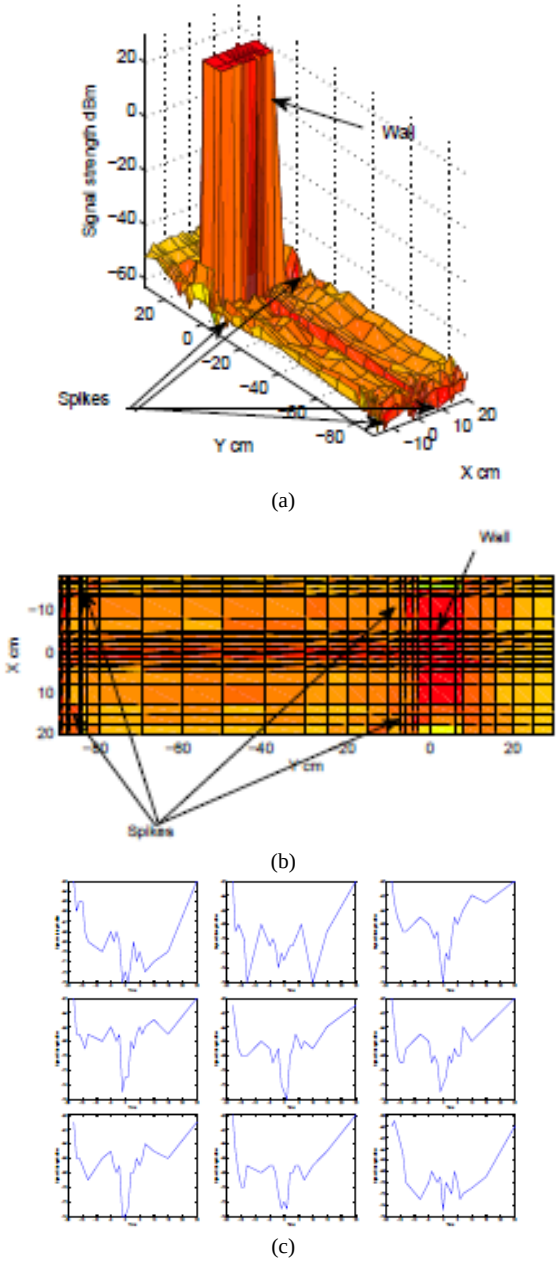
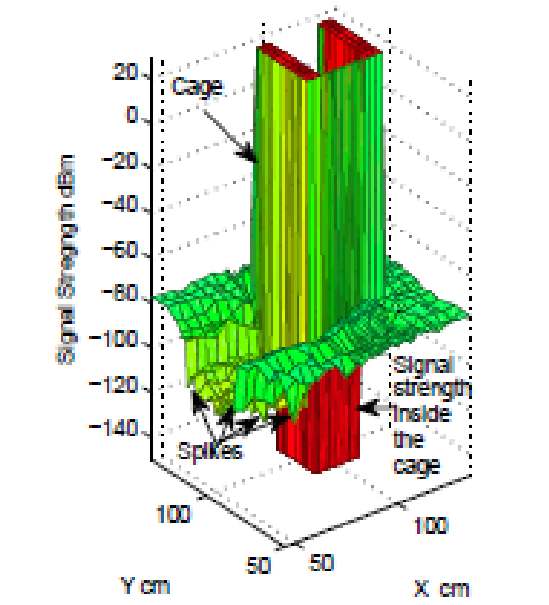
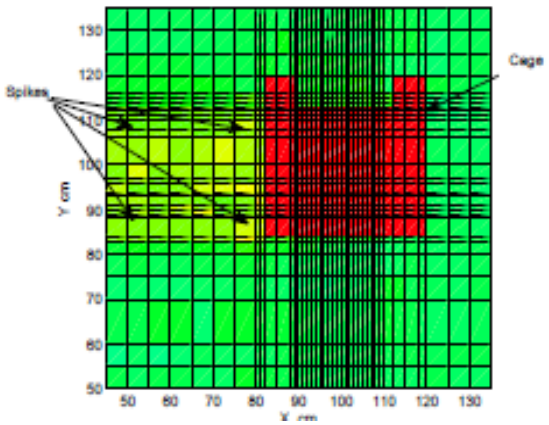


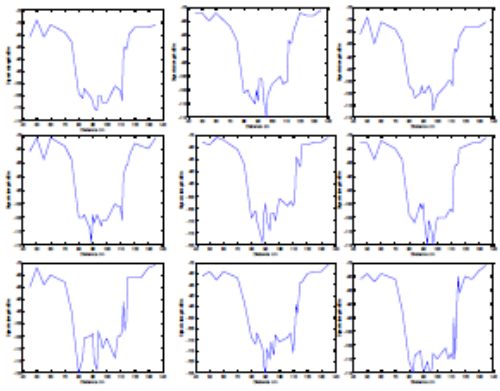
Fig. 8. Copper Wall Obstacle (a) RF-SS Measurements (b) Top view of Fig. 7(a), and (c) Multiple Signal Shapes.



(a)



(b)



(c)

Fig. 9. Copper Cage Obstacle (a) RF-SS Measurements (b) Top view of Fig. 8(a), and (c) Multiple Signal Shapes.

As a result of LOS loss between the transceivers, the SS dropped down and produced poor conductive connectivity. It is shown in Fig. 9(b), the top view of Fig. 9(a). The SS improves as the transceivers retain the LOS and become more conductive, approximating the simulations as depicted in Fig. 4(a). In Fig. 9(b), the backside of the cage, the SS reaches the maximum conductivity as the transceiver moves further. The results depicted in Fig. 9(a) override the results illustrated in Fig. 5(a) by a value of -5dB, resulting from different surrounding electromagnetic sources. Fig. 9(c) illustrates the signature and the obstacle impact on RF-SS, showing different signal shapes for varied trajectories.

V. Validation of Physical and Simulation Results

The simulation and physical results comparison and validation of the RF obstacle discussed in previous sections are presented. For accuracy and comparison, different signal shapes of the obstacles are plotted in the same graph. Additionally, the effect of various electromagnetic sources on the physical signal shapes is detectable in the signal shapes.

A. Validation of Wall Results

Using the setup of Fig. 7, multiple wall physical experiments are conducted to demonstrate the simulation results. The experiments are conducted in an environment with fewer interference sources. The transceivers are moved in a bounded area of 2 m2 around a centered wall obstacle in all directions. The resulting signal shapes for the simulation (black) and physical (red and blue) are depicted in Fig. 10. The signal power difference between the black and red signals ranges from -2 to -8 dBm, while it was above -15 dBm between black and blue signals due to interference source existence.

B. Validation of Cage Results

Numerous cage physical experiments are conducted to demonstrate the simulation results. First, the experiments are conducted in an environment with fewer interference sources. Then, the transceivers are moved in a bounded area of 2 m2 around a centered cage obstacle in all directions.

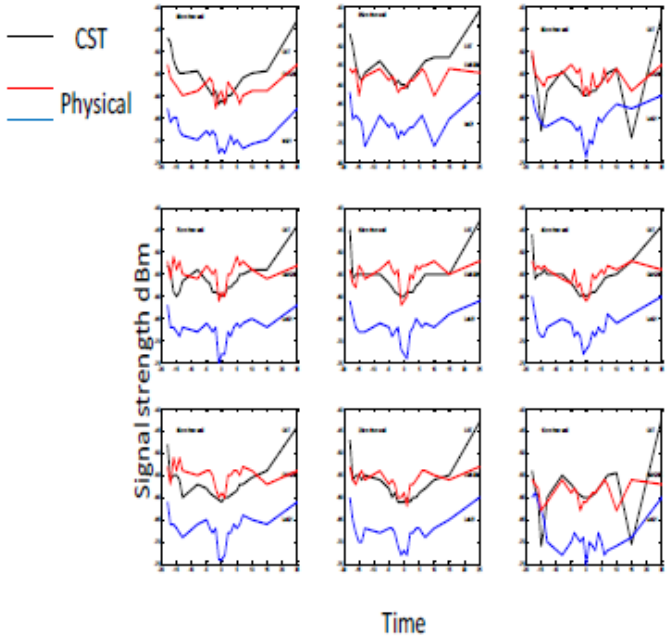


Fig. 10. Wall Simulation and Physical Results Comparison.

The resulting signal shapes for the simulation (black) and physical (red and blue) are depicted in Fig. 11. The signal power difference between the black and red signals ranges from -3 to -20 dBm, while it was above -20 dBm between black and blue signals due to interference source existence.

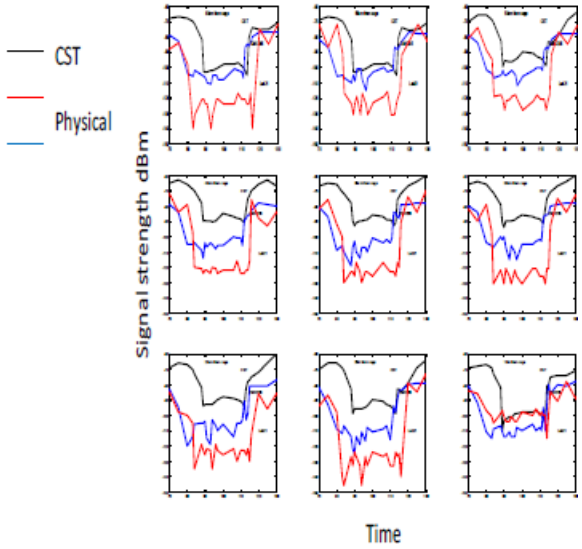


Fig. 11. Cage Obstacle Physical Results Comparison.

VI. Obstacle Parameterization

Radio SS propagation is a complicated process. In Sections I and II, we explained that SS is a function of different parameters. In addition, the power of the received SS is a function of how far the transmitter is, obstacles effects, and multipath occurrences such as reflections and refractions [2,15].

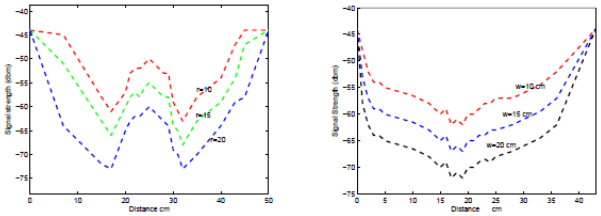


Fig. 12. Signal shapes for different (a) Different Radiuses (b) Wall widths.

A. Wall Parameterization

Different RF walls obstacles are used in our experiments with varying sizes of width w. As a result, the wall signal shapes are almost preserved in the "U" shape. The signal shapes are scaled and stretched as the width of the wall increases. Table I shows the average signal strength in dBm on the robot's trajectories for different wall widths. As the wall width increases by 1 cm, the signal strength average alongside the robot path decreases by −1.2 dBm, as shown in Fig. 12(b).

TABLE I. Average SS on the Robot Trajectory

Cylinder radius r	10 cm	15cm	20 cm
Average SS(dBm)	-53	-59	-64

VII. RFMR Method based on HMM

The RFMR method is summarized in the significant steps shown in the diagram of Fig. 13. Foremost, in Fig. 14, the measurement vector acquired through multiple robot motion paths is split into various segments (small components) of comparable lengths. Afterward, features are extracted in the frequency domain by applying Fast Fourier Transform (FFT) on every element of the segmented signal. The features components extracted are written to vectors, and then we used a subset of the vectors of features for the training purpose of the created model and the remaining vectors used for model testing. Next, the training subset is clustered to generate observation sequences using the K-mean clustering algorithm [31]. Then, three HMMs models are trained using the generated observation sequences. Each HMM model that contains five states is assigned to each obstacle type. The five states correspond to 5 small segments produced on the robot-specific trajectory. As illustrated, we trained each HMM model using a specific set of observation sequences. Finally, the training set of features is used to train classification models. Accordingly, we accomplished the RFMR method results [28]. Consequently, the results were used by the robot PMCA algorithm that uses the trained HMM results. As a result, we accomplished proactive connectivity [30].

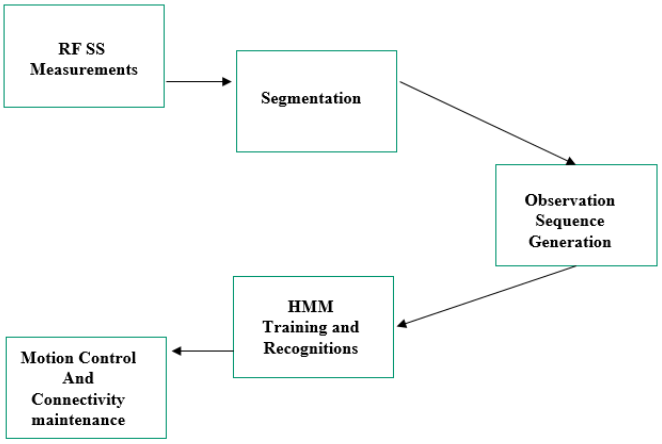


Fig. 13. RFMR Method Block Diagram.

The robot's movement is a sequential event, and our goal is to classify the robot movement in sequential order of the segments. Furthermore, there is a strong analogy between the RFMR method based on HMM results and the word recognition via speech patterns [32]. Therefore, using HMM offers a more spontaneous methodology for RF shadows classification. Naturally, it breakdowns the measurements on the robot's pathway to approximate components comparable to how they were created. However, the HMM method offers a simple technique for classifying segments subset on the robot's path as it moves through an obstacle shadow rather than categorizing the pieces after the obstacle shadow. The HMMs and their application in RFMR are discussed in the following sections. Hence, HMM is a method for stochastic events of a model. Clearly, A model λ consists of several states Q, observations B corresponding probabilities, and transitions between states probabilities [11]. Therefore, specified a sequence of observations, O, and λ as a model, one

can obtain P(O|λ). Fundamentally, it is the model representation of the event, and it could be a good or bad representation. To classify data using HMM, we need to create a model λ(j), j=1, …, m_o for each class, where m_o denotes obstacle types number. Then, we must calculate P(O|λ(j)) corresponding to each obstacle type available. Finally, the model with the highest probability is allocated to a novel observation O. Therefore, the obstacle type membership is given to O.

A. Feature Extraction based on SS Measurement Segmentation

The collected SS measurement vector through the robot moves on the trajectory, *lth*, is $\beta_l^{(j)} = [S_l^{(j)}(1), S_l^{(j)}(2), \ldots S_l^{(j)}(Nm)]^T$, For the *j*th obstacle type, N_m is the SS measurements number on the *l*th path. Then, as in Fig. 14, each $\beta_l^{(j)}$ is segmented into five segments represented as $\alpha_{l.u}^{(j)} = [S_l^{(j)}((u-1)+1) \ldots S_l^{(j)}(5u)]^T, where\ u = 1,2,\ldots,5,$. Subsequently, a measurement segment $\alpha_{l.u}^{(j)}$ is transformed to the using FFT. Results are represented as $r_{l.u}^{(j)} = FFT(\alpha_{l.u}^{(j)}, N_{FFT})$, N_{FFT} shows the points number in the FFT results. The first ten elements in the FFT result $r_{l.u}^{(j)}$ are denoted as the feature vector $\gamma_{l.u}^{(j)} = [r_{l.u}^{(j)}(1)\ r_{l.u}^{(j)}(2) \ldots r_{l.u}^{(j)}(10)]^T$ Of the measurement corresponding to the *l*th trajectory and *j*th obstacle type. Once each segment is (j) transferred into frequency space, the feature vector $r_{l.u}^{(j)}$ is clustered using the K-means clustering algorithm [35].

Then, the HMM uses these binned segments to classify the obstacle shadow based on the probabilistic sequence of segments. Our numerical experiments tried different training sets to examine their effect on the recognition rate. We found that the recognition rate is affected positively by the size increase of the training sets. Data were randomly split into training and testing sets to verify the HMM classifier [35]. We randomly select 60% of the measurement vectors into the training set S_{train}^c which is used for c clustering and training, and the rest constitutes the testing set S_{test}^c.

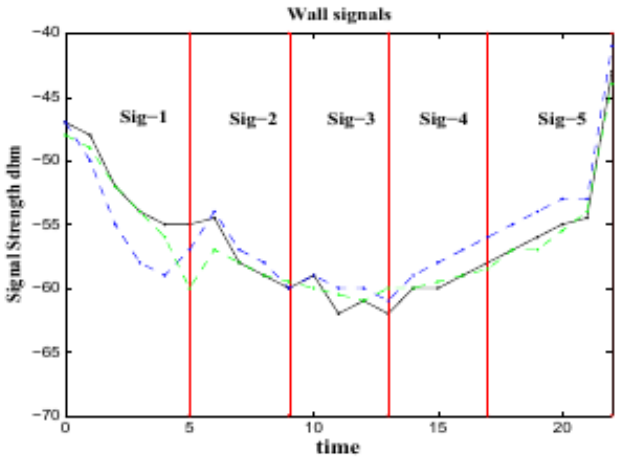


Fig. 14. One Segmented Wall Signal.

B. The Generation of Observation Sequences

The collected vectors $\gamma_{l.u}^{(j)}$ in the model training set S_{train}^c are split into *G* different clusters using the k-means algorithm. The *G* clusters are arranged as $D_1, D_2, \cdots, D_G$. Therefore, we can minimize the within-cluster sum of squares (WCSS). Equation (3) presents the *k*-means algorithm stages, and it is written as.

$$\arg min_{D1,\cdots,DG} \sum_{g=1}^{G} \sum_{\beta_l^{(j)} \in S_{train}^c, \gamma_{l.u}^{(j)} \in Dg} \| \gamma_{l.u}^{(j)} - \mu_g \|^2 \qquad (3)$$

where the parameter μ_a represents the centroid of D_a, i.e., the points mean in D_g, $\| \gamma_{l.u}^{(j)} - \mu_g \|^2$ represents the vector $\gamma_{l.u}^{(j)}$ and μ_a distance separation. After the k-means clustering algorithm generates D_a and μ_a, it allocates observation symbols to the feature vectors. As a result, the observation sequences for HMM training and test sets are produced. Initially, The $C = \{C_1, \cdots, C_G\}$ with C_g as the *g*th symbol signifies the symbol set of HMM observations [34]. The $C_{l.u}^{(j)}$ symbol conforming to the data segment $\gamma_{l.u}^{(j)}$ is allocated to the value C_g if $\| \gamma_{l.u}^{(j)} - \mu_g \|^2$ has the minimum value in the set, $g \in \{1, 2, \cdots, G\}$. Explicitly, $C_{l.u}^{(j)}$ is allocated to symbol C_g when μ_g is the closer centroid to the feature vector $\gamma_{l.u}^{(j)}$. The $C_{l.u}^{(j)}$ vectors of the *l*th trajectory segments are concatenated to form the vector $C_l^{(j)} = [C_{l.1}^{(j)} \ldots C_{l.5}^{(j)}]^T$ [34]. The resulting vector $C_l^{(j)}$ is the observation sequence corresponding to the measurement vector $\beta_l^{(j)}$. The HMM training set S_{train}^{HMM} conforming vector $\beta_l^{(j)}$ contains the observation sequence $C_l^{(j)}$ is included in the training set S_{train}^c; otherwise, it is in the test set S_{test}^{HMM}.

In conclusion, applying the mentioned procedure, for the trajectory, *lth*, in the existence of type *j* obstacle, *the* vector $\beta_l^{(j)}$ is segmented into segments $\alpha_{l.u}^{(j)}$, *u* = 1, 2, 3, 4, 5. Consequentially, each $\alpha_{l.u}^{(j)}$ is converted by FFT, and the FFT result is symbolized by $r_{l.u}^{(j)}$. The first ten elements in $r_{l.u}^{(j)}$ are selected to form the feature vector $\gamma_{l.u}^{(j)}$. The feature vectors are clustered using the *k*-means algorithm to generate *G* clusters, $D_1, \cdots, D_G$, and the corresponding cluster centroids $\mu_1, \ldots, \mu_G$.

The individual segment feature vector $\gamma_{l.u}^{(j)}$ is consigned with a symbol $C_{l.u}^{(j)} \in C$ by the parameters of the cluster [35]. Then, the observation sequence $C_l^{(j)}$ is created by concatenating the $C_{l.u}^{(j)}$ vector [34]. Finally, the observation sequence $C_l^{(j)}$ is created from each trajectory $\beta_l^{(j)}$ vector and ready for training or testing HMMs models [34, 35].

VIII. RFMR BASED HMM NUMERICAL RESULTS

The HMM training set S_{test}^{HMM} resulted from the previous section will be used for training HMMS models. Three HMMs, models, *λ(j)*, with *j* = 1, 2, 3, conforming wall, cage, and cylinder obstacles, are trained for classification.

Assumed an observation sequence $C_l^{(j)}$, is comprised of numerous observation symbols. Next, given the HMM λ(p), the P (Cl | λ(p)) is a conditional probability of $C_l^{(j)}$. For the classification process, the P (Cl | λ(p)), can be calculated for the parameter p = 1, 2, 3. When the value $p = \hat{p}$, the maximum probability $P(C_l| \lambda_{(p)})$ is achieved, and the obstacle of type $\hat{p}$ is the prediction of RFMR method results.

Moreover, if a transceiver approaches an obstacle while the complete observation sequence was not existing, the first few observations for the classification.is found in the observation sequence of a variable-length vector $C_l^{(j)}$.

A. Examining Cylinder Obstacle of different Radius

The total measurement vectors used are 535. A set of 321 vectors is for the training, and the rest is for testing. These measurement vectors contain data from three different cylinder radiuses, 10 cm, 15 cm, and 20 cm, with a height of 30 cm. The confusion matrix (CM) that signifies the RFMR results is depicted in Tables II, III, and IV, each row of the CM denotes the predicted class. Table II establishes the CM using the first two segments of observation sequences; a percentage of 88% was the classification rate attained. Table III reflects the CM using the first three segments, and the rate was 95%. Finally, a rate of 100% was reached using four segments as presented in Table IV. The results are improved for the HMM classifier as the number of segments increases, and consequently, the rates become reliable.

B. Examining Wall Obstacle of Various Dimensions

The total measurement vectors used are 455. A subset of two hundred seventy-three vectors is for the training, and the rest is for testing. The measurement vectors contain data for 7 × 30 × 30 cm^3, 10 × 30 × 30 cm^3, and 15 × 30 × 30 cm^3 wall dimensions. The CM of the RFMR results is shown in Tables V, VI, and VII. Table V demonstrates the CM of RFMR of wall measurement vectors based on the first two segments, and the rate was 70%.

In comparison, Table VI presents the CM with the first three segments, and the rate was 77%. Finally, Table VII validates the CM of RFMR results based on the first four segments; the success rate was 93%. The results are improved for the HMM classifier as the number of segments increases, and consequently, the rates become excellent.

TABLE II. CM OF RFMR FOR CYLINDER USING 2 OBSERVATIONS

Cylinder radius r	10 cm	15cm	20 cm
10 cm	1	0.0	0.14
15 cm	0.0	1	0.22
20 cm	0.0	0.0	0.64

TABLE III. CM OF RFMR FOR CYLINDER USING 3 OBSERVATIONS

Cylinder radius r	10 cm	15cm	20 cm
10 cm	0.86	0.0	0.0
15 cm	0.14	1	0.0
20 cm	0.0	0.0	1

TABLE IV. CM OF RFMR FOR CYLINDER USING 4 OBSERVATIONS

Cylinder radius r	10 cm	15cm	20 cm
10 cm	1	0.0	0.0
15 cm	0.0	1	0.0
20 cm	0.0	0.0	1

TABLE V. CM OF RFMR FOR WALL USING 2 OBSERVATIONS

Wall width (w)	7 cm	10 cm	15 cm
7 cm	0.85	0.46	0.0
10 cm	0.15	0.54	0.31
15 cm	0.0	0.0	0.69

TABLE VI. CM OF RFMR FOR WALL USING 3 OBSERVATIONS

Wall width (w)	7 cm	10 cm	15 cm
7 cm	0.87	0.44	0.0
10 cm	0.13	0.56	0.16
15 cm	0.0	0.0	0.84

TABLE VII. CM OF RFMR FOR WALL USING 4 OBSERVATIONS

Wall width (w)	7 cm	10 cm	15 cm
7 cm	0.95	0.08	0.0
10 cm	0.05	0.92	0.12
15 cm	0.0	0.0	0.88

TABLE VIII. CM OF RFMR FOR ALL OBSTACLES USING 2 OBSERVATIONS

Different Obstacle	Cage 30 cm^3	Wall 10 cm	Wall 15 cm	Cylinder 10 cm	Cylinder 15 cm
Cage 30 cm^3	1	0.0	0.0	0.0	0.0
Wall w = 10 cm	0.0	1	0.44	0.00	0.0
Wall w = 15 cm	0.0	0.0	0.56	0.0	0.0
Cylinder r = 10 cm	0.0	0.0	0.00	0.80	0.0
Cylinder r = 15 cm	0.0	0.0	0.0	0.20	1

C. Examining Walls, Cages and Cylinders Obstacle of Different Sizes

In the experiment that combines three different size obstacles, the total measurement vectors used are 825. Four hundred ninety-five vectors are the training set and the rest for testing. Tables VIII, IX, and X illustrate the CM of the RFMR results for all combined obstacles observation vectors where the predicted class is expressed by CM rows approximated to the actual class. Table VIII establishes the CM of RFMR results using the first two segments, and the classification rate is 87%. When increasing the segment number to three, the success rate was 89%, as shown in Table IX. Ultimately, the classification rate increases and reaches 92% as the segments number increased to four and above, as illustrated in Table X.

TABLE IX. CM OF RFMR ALL OBSTACLES USING 3 OBSERVATIONS

Cage 30 cm^3	1	0.0	0.0	0.0	0.0
Wall w = 10 cm	0.0	0.34	0.0	0.00	0.0
Wall w = 15 cm	0.0	0.66	0.44	1	0.0
Cylinder r = 10 cm	0.0	0.0	0.56	0.80	0.0
Cylinder r = 15 cm	0.0	0.0	0.0	0.20	1

TABLE X. CM OF RFMR FOR ALL OBSTACLES USING 4 OBSERVATIONS

Different Obstacle	**Cage 30 cm^3**	**Wall 10 cm**	**Wall 15 cm**	**Cylinder 10 cm**	**Cylinder 15 cm**
Cage 30 cm^3	1	0.0	0.0	0.0	0.0
Wall w = 10 cm	0.0	1	0.40	0.00	0.0
Wall w = 15 cm	0.0	0.0	0.60	0.0	0.0
Cylinder r = 10 cm	0.0	0.0	0.0	1	0.0
Cylinder r = 15 cm	0.0	0.0	0.0	0.00	1

In conclusion, the results are improved for the HMM classifier as the number of observation segments increases, and consequently, the rates of successful classification become promising and outstanding. Therefore, it proves that the proposed methods are reliable for the best classification rates, thus, achieving proactive robot control in the field.

IX. PROACTIVE MOTION CONTROL ALGORITHM (PMCA) FOR PRESERVING CONNECTIVITY

The developing application of mobile robotics networks has produced the control motion concept of mobile nodes communication, so nodes can preserve connectivity while finishing their tasks in the field [8, 34]. However, the control motion techniques require deep exploration and the creation of more reliable algorithms in the robotic field [7]. When the SS drops down in the field, and a robot loses the collaborating robots in the swarm, it starts preserving connectivity through the movement control algorithm, which assists the robot in reaching a location in the field, where it can gain coverage communicate with other team members. Depending on the results of the RFMR method, the proposed PCMA algorithm decides either to continue the current trajectory or backward movement until it retains reliable SS. The PCMA control decision is mainly based on the information learned from the obstacle shadow recognition.

The proactive control motion algorithm has two choices to achieve its motion control of mobile robots. As mentioned earlier, the control choices are based on the information learned from RFMA results. Firstly, the PMCA can continue to move robots in the current trajectory across the obstacle until the robots preserve communication successfully. The second control choice is moving the robot backward and computing a 4-D gradient based on SS to define the strong SS direction and then communicate with the team [33]. In summary, algorithm one and the flowchart of PCMA in Fig. 15 illustrate the actual steps to control the mobile robot motion to preserve communication.

Algorithm 1: Proactive Motion Control Algorithm (PMCA)

```
1: Input: RFMR result.
2: Output: Maintaining connectivity of mobile rebots.
3: Get RFMRRecognitionResults()
4:   if (Obstacle type and size are estimated) then
5:      if Segments length ≥ (estimated size/2) then
6:            MoveCurrentPath()
7:            GradientDecsentAlgorithm()
8:      else
9:            MoveBack()
10:           GetStrongSignalPos()
11:           GradientDecsentAlgorithm()
12:     end if
13:  else
14:        Moveback()
15:        GetStrongSignalPos()
16:        GradientDecsentAlgorithm()
17: end if
18: MaintainConnextivity()
```

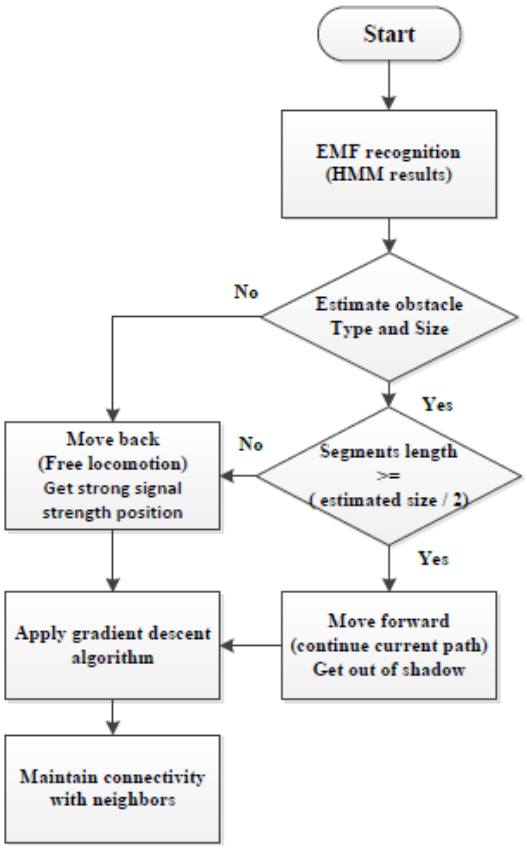


Fig. 15. The PMCA Flow Chart.

X. Gradient-based on the Proactive Control Algorithm

The gradient computation process in the field is based on measuring SS between to mobile transceivers that are at d distance apart as in Fig. 20, where the SS is measured at the receiver side. The signal $S_l^{(j)}(t)$ at time t of the lth trajectory around an obstacle type *j* is calculated according to Equation (2). When mobile robots move and retain LOS, $S_l^{(j)}(t)$ is stable and preserves robots connectivity. In contrast, the SS dropped as a conductive obstacle blocks the moving robots [34].

Accordingly, the signal measurements, $S_l^{(j)}(t)$, collected through the robot's motion at the position $(x_t^{(i)}, y_t^{(i)})$, i = 1, 2 at time *t*. Next, the gradient is calculated for a specific robot trajectory [34]. For the *l*th trajectory, the gradient vector can be expressed as

$$\nabla S_l^{(j)}(t) = \left[\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(1)}} \frac{\partial s_l^{(j)}(t)}{\partial y_t^{(1)}} \frac{\partial s_l^{(j)}(t)}{\partial x_t^{(2)}} \frac{\partial s_l^{(j)}(t)}{\partial y_t^{(2)}}\right]^T \quad (4)$$

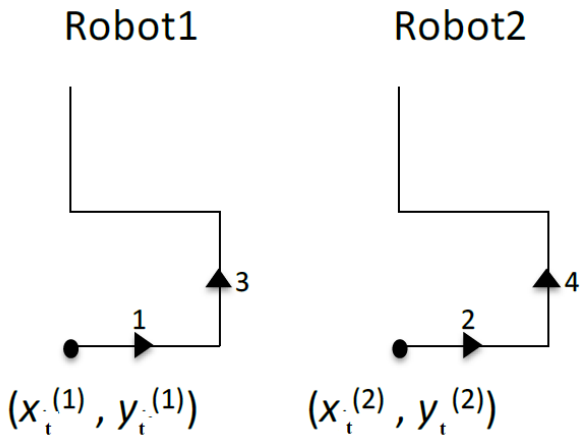


Fig. 16. The Transmitter and Receiver Robot Stepwise Trajectory.

The SS gradient is calculated following the of Fig. 16, the initial position at time *t* for robot one and robot two are $(x_t^{(1)}, y_t^{(1)})$ and $x_t^{(2)}, y_t^{(2)})$. As explained in the method below, we assumed that one robot is moving while the other stays still to compute the gradient. As pointed out in Fig. 16, during time t and t+1, robot one moves along trajectory segment 1, so $x_{t+1}^{(1)} = x_t^{(1)} + \Delta x$, $y_{t+1}^{(1)} = y_t^{(1)}$, $x_{t+1}^{(2)} = x_t^{(2)}$, $y_{t+1}^{(2)} = y_t^{(2)}$, and the gradient element $\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(1)}}$ is computed a.

$$\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(1)}} \approx \frac{\partial s_l^{(j)}(t)}{\partial x_t^{(1)}} = \frac{s_l^{(j)}(t+1) - s_l^{(j)}(t)}{\Delta x} \quad (5)$$

Throughout time t + 1 and t + 2, robot two moves along segment 2, so $x_{t+2}^{(1)} = x_{t+1}^{(1)}$, $y_{t+2}^{(1)} = y_{t+1}^{(1)}$, $x_{t+2}^{(2)} = x_{t+1}^{(2)} + \Delta x$, $y_{t+2}^{(2)} = y_{t+1}^{(2)}$ And the gradient element $\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(2)}}$ is computed as.

$$\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(2)}} \approx \frac{\partial s_l^{(j)}(t+1)}{\partial x_{t+1}^{(2)}} = \frac{s_l^{(j)}(t+2) - s_l^{(j)}(t+1)}{\Delta x} \quad (6)$$

In time t + 2 and t + 3, robot one moves along segment 3, so $x_{t+3}^{(1)} = x_{t+2}^{(1)}$, $y_{t+3}^{(1)} = y_{t+2}^{(1)} + \Delta y$, $x_{t+3}^{(2)} = x_{t+2}^{(2)}$, $y_{t+2}^{(2)} = y_{t+1}^{(2)}$, and the gradient element $\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(2)}}$ is computed as.

$$\frac{\partial s_l^{(j)}(t)}{\partial y_t^{(1)}} \approx \frac{\partial s_l^{(j)}(t+2)}{\partial y_{t+2}^{(2)}} = \frac{s_l^{(j)}(t+3) - s_l^{(j)}(t+2)}{\Delta y} \quad (7)$$

In time t + 3 and t + 4, robot two moves along trajectory segment 4, so $x_{t+4}^{(1)} = x_{t+3}^{(1)}$, $y_{t+4}^{(1)} = y_{t+3}^{(1)}$, $x_{t+3}^{(2)} = x_{t+2}^{(2)}$, $y_{t+4}^{(2)} = y_{t+3}^{(2)} + \Delta y$, and the gradient element $\frac{\partial s_l^{(j)}(t)}{\partial x_t^{(2)}}$ is computed as.

$$\frac{\partial s_l^{(j)}(t)}{\partial y_t^{(2)}} \approx \frac{\partial s_l^{(j)}(t+3)}{\partial y_{t+3}^{(2)}} = \frac{s_l^{(j)}(t+4) - s_l^{(j)}(t+3)}{\Delta y} \quad (8)$$

As shown in Fig. 17, arrows indicate the gradient direction, and yellow grids illustrate reliable SS due to LOS existence between the communicated transceivers. The gradient strength and direction depend on the robot's location concerning the obstacle position in the field. For example, if one robot is surrounded inside the cage, the gradient drops down as green boxes indicate. Accordingly, any movement for the outer robot did not improve the SS for communication. However, when the robot in the cage changes position, the SS improves enough for communication, as in Fig. 17.

The scenario of Fig. 18 illustrates an obstacle and two robots in the simulation field. One robot moves in a stepwise trajectory, and the other is stays still. Consequently, the gradient is scattered when no LOS exists and does not contain useful information due to obstacle shadow. However, the gradient improved as the LOS became clear.

The scenario of Fig. 19 illustrates an obstacle and two robots in the simulation field. One robot stays still closer to the obstacle corner while the robot moves through a stepwise trajectory. The gradient is scattered when no LOS exists and does not contain useful information due to obstacle shadow. However, the gradient improved as the LOS became clear. In summary, the gradient helps find the right direction of the partner, as illustrated by Fig. 18 and Fig. 19.

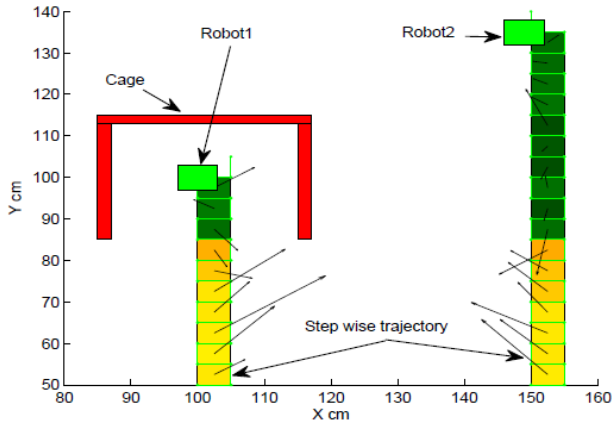


Fig. 17. One Robot is trapped in the Cage, the other Moves in a Stepwise Trajectory.

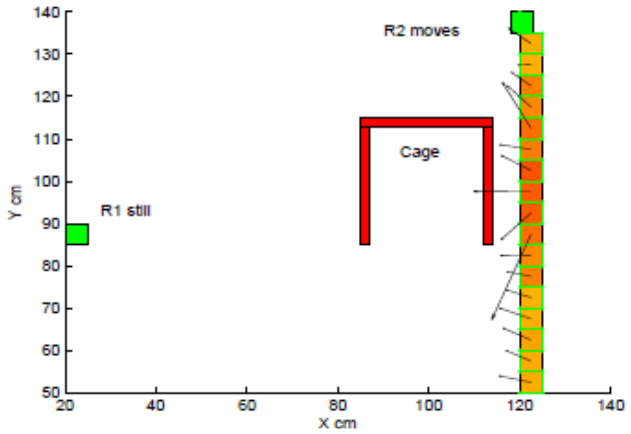


Fig. 18. One Robot Moves in a Stepwise Trajectory, and the other Stays Still.

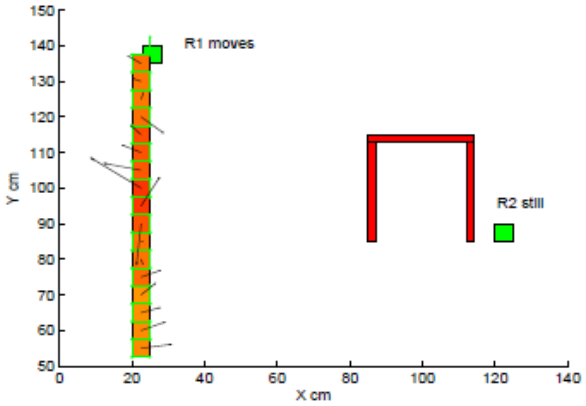


Fig. 19. One Robot Moves in a Stepwise Trajectory, and the other is Close to the Obstacle.

The scenario of Fig. 20 illustrates an obstacle and two robots in the simulation field. It shows different trajectories for one robot moving straight and the other through a stepwise course. The gradient is scattered when no LOS exists and does not contain useful information due to obstacle shadow. However, the gradient improved as the LOS became clear. The gradient is computed according to Section V's equations (5) and (7). The gradient helps find the right direction of the other robot and preserve connectivity.

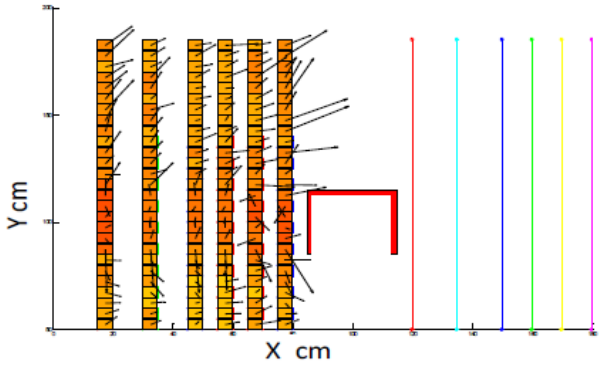


Fig. 20. Different Robots Trajectory around the Cage.

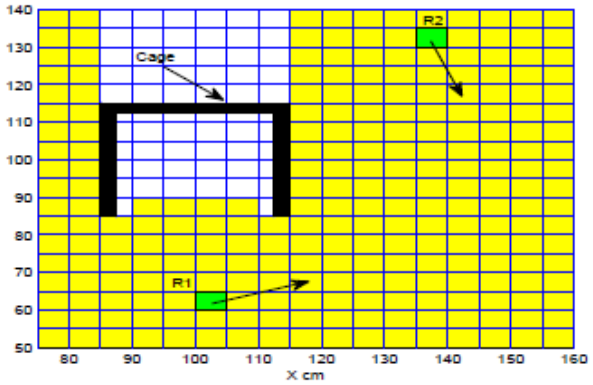


Fig. 21. Configuration Space of Two Robots.

XI. The Experimental Results of the Gradient Algorithm

Section III explains that the experimental robot field is divided into equal grids. Then, two robots move, measure SS and compute the gradient for any two grids in the area marked yellow in Fig. 21. Next, we created a database containing the robot's position, measured SS, and calculated the gradient for any two grids (robot's location) at time t. The main steps of the algorithm are illustrated in Fig. 22. For example, at time $t = 0$, two robots are placed at $x_0^{(1)} = 20$; $y_0^{(1)} = 5$ and robot 2 starts at $x_0^{(2)} = 40$; $y_0^{(2)} = 5$ in the field. Fig. 23 depicts the trajectories of the robots resulting from running the gradient algorithm, and results confirm that the algorithm helps evade obstacles' shadows and preserve communications. Fig. 24 depicts another scenario where the robots are placed in front of the obstacle at ($x_0^{(1)} = 18$; $y_0^{(1)} = 14$) and the other robot at ($x_0^{(2)} = 34$; $y_0^{(2)} = 16$). The gradient algorithm exhibits promising results to preserve communication between mobile robots.

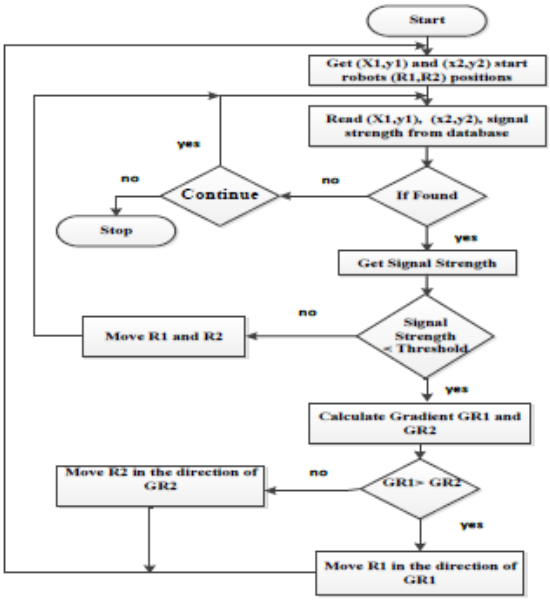


Fig. 22. Gradient Algorithm Flowchart.

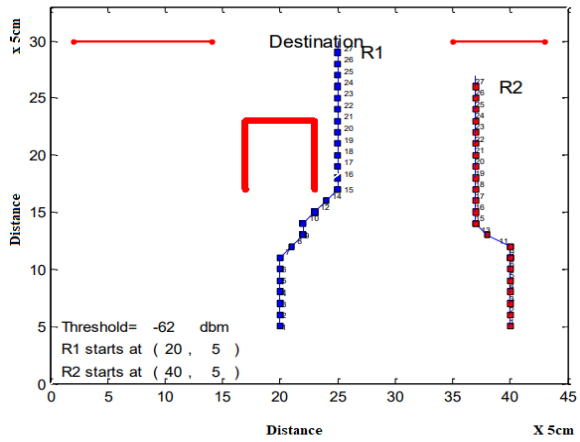


Fig. 23. Robots Start at $x_0^{(1)}$= 20; $y_0^{(1)}$ = 5 and $x_0^{(2)}$= 40; $y_0^{(2)}$= 5 at Time t=0.

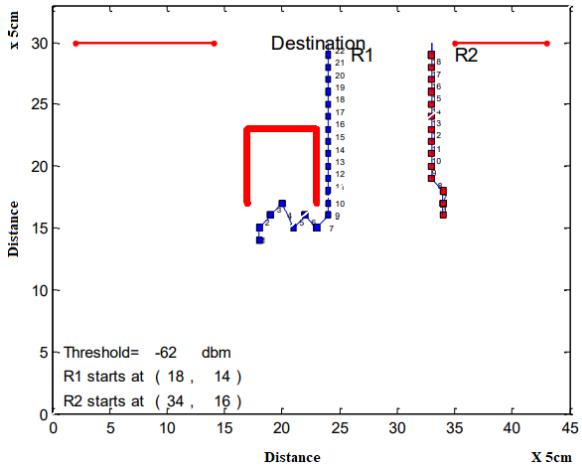


Fig. 24. Robots Start at $x_0^{(1)}$) = 18; $y_0^{(1)}$= 14 and $x_0^{(2)}$= 34; $y_0^{(2)}$= 16 at time t=0.

Another scenario is depicted in Fig. 25, where one robot faces the obstacle at $x_0^{(1)}$ = 17; $y_0^{(1)}$= 5 and the other at $x_0^{(2)}$ = 38; $y_0^{(2)}$= 5 at time t = 0. The results confirm the algorithm's success in preserving communication.

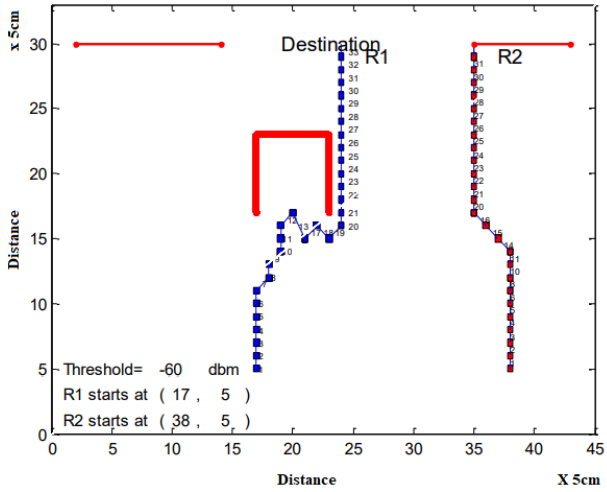


Fig. 25. Robots Start at $x_0^{(1)}$) = 17; $y_0^{(1)}$= 5 and $x_0^{(2)}$= 38; $y_0^{(2)}$= 5 at time t=0.

XII. The Experimental Results of the PMCA

The HMM results of the RFMR method demonstrate the detection of obstacles on the robotic path and determine the type of distance from the robot path and the size of the approximate obstacle. PMCA uses HMM results to encourage the robot to continue moving through the current trajectory based on the length of the segments covered by the robot. If the segment's length is equal to half or greater than the estimated obstacle size, then the robot continues forward. Otherwise, the robot stops and returns to a position with robust signal strength, as shown in the scene in Fig. 26. Afterward, the robot runs a gradient algorithm to determine the strong SS direction. After that the robot moves in the gradient trend and re-establishes connectivity, as shown in the scene in Fig. 27. Algorithm 1 and Fig. 15 illustrate the PMCA mechanism.

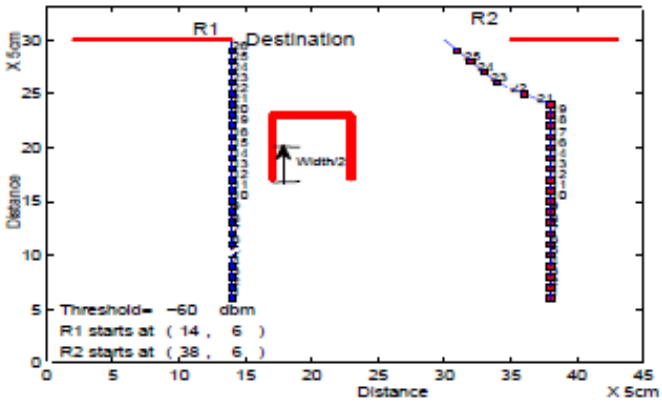


Fig. 26. PCMA uses Three Observations.

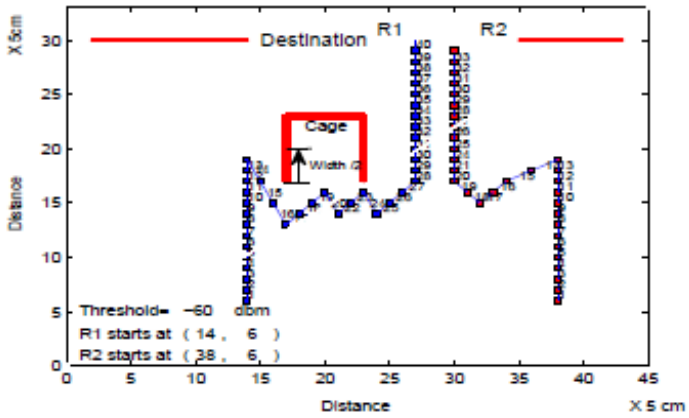


Fig. 27. PCMA uses Two Segments.

XIII. Conclusion

The article introduces Radio Frequency Mapping Environment Recognition (RFMR), gradient, and proactive robot motion control algorithms (PMCA). Thus, we conducted many simulations and physical experiments to assess the proposed method's performance. Consequently, this work presents promising solutions and becomes a competitive alternative for the routing and maintaining broken links problems in robot networks. Furthermore, extensive

simulation and physical experiments will be conducted to validate the recognition of different RF obstacles. Also, obstacle parameterization and generalization approaches will be addressed in future studies.

REFERENCES

[1] Y. Wu , X. Ren, H. Zhou , Y. Wang. and X. Yi. "A Survey on Multi-Robot Coordination in Electromagnetic Adversarial Environment: Challenges and Techniques," in IEEE Access, 2020, vol. 8, pp. 53484-53497.

[2] M. Lindhe, H. Johansson, A. Bicchi. "An experimental study of exploiting multipath fading for robot communication," in Proceeding of Robotics: Science and Systems,2007, Atlanta.

[3] R. M. Voyles, J. Bae, A. Larson. and M. Ayad. "Wireless video sensor network for sparse, resource-constrained, multi-robot teams," in Journal of Intelligent Service Robots,2009.vol. 2, no. 4, pp.235--246.

[4] J. Bae. and R. M. Voyles. "Wireless video sensor network from a team of urban search and rescue robots," in International Conference on Wireless Networks, 2006.

[5] N. Michael, M. M. Zavlanos, V. Kumar and G. J.Pappas." Maintaining connectivity in mobile robot networks," in International Symposium on Experimental Robotics. July 2009, Athena, Greece.

[6] D. Hahnel, B. Ferris. and D. Fox. "Gaussian processes for signal strength-based location estimation," in Proceeding of Robotics: Science and Systems, 2006, Philadelphia, PA.

[7] V. Loscri, E. Natalizio. and C. Costanzo. "Simulations of the impact of controlled mobility for routing protocols," in EURASIP Journal on Wireless Communications and Networking, July 2010.

[8] A. Purohit, S. Zheng, F. Mokaya, and P. Zhang." Sensorfly: Controlled-mobile sensing platform for indoor emergency response applications," in Information Processing in Sensor Networks (IPSN), 10th International Conference, 2011.

[9] A. Ghaffarkhah, and Y. Mostofi. "Path Planning for Networked Robotic Surveillance," in Signal Processing, IEEE Transactions, July 2012, vol. 60, no. 7, pp. 3560-3575.

[10] T. Kanungo, M. David., S. Nathan, Netanyahu, D. Christine., S. Ruth, and Y. Angela. "An efficient k-means clustering algorithm: Analysis and implementation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, July 2002, vol. 24, no. 7, pp. 881--892.

[11] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition, "Feb 1989, vol.~77, pp. 257 --286.

[12] He. Zhihai and Wu. Dapeng. “Resource allocation and performance analysis of wireless video sensors,” in IEEE Transactions on Circuits and Systems for Video Technology, May 2006,vol.~16.

[13] M. Ayad, and R.Voyles. "Physical Link Maintenance and Logical Message Routing Integration for Robotic Network Connectivity," in Vehicular Technology Conference, 2019, IEEE. Hawaii.

[14] J. Sweeney, T. J. Brunette, Y. Yang, and R. Grupen. "Coordinated teams of reactive mobile platforms"' in International. Conference on Robotics and Automation (ICRA). IEEE, 2002, vol.~1, pp. 299--304.

[15] J. Fink, and V. Kumar. "Online methods for radio signal mapping with mobile robots network," in International Conference on Robotics and Automation(ICRA, IEEE, 2010.

[16] J. N. Twigg, J. R. Fink, P. L. Yu, and B. M. Sadler. "RSS gradient-assisted frontier exploration and radio source localization," in International. Conference on Robotics and Automation(ICRA). IEEE, May 2012, pp. 889--895.

[17] Yi. Sun, X. Jizhong, Li. Xiaohai, and F. Cabrera-Mora. "Adaptive source localization by a mobile robot using signal power a gradient in sensor networks," in IEEE Global Telecommunications Conference, 2008, pp. 1 --5.

[18] Yan. Yuan, and Y. Mostofi. "Co-optimization of communication and motion planning of a robotic operation in fading environments," in Signals, Systems, and Computers (ASILOMAR) Conference Record of the Forty Fifth Asilomar Conference, Nov. 2011, pp. 1455--1460.

[19] J. Fink, V. Kumar, N. Michael, and A. Kushleyev." Experimental characterization of radio signal propagation in indoor environments with application to estimation and control," in IEEE International Conference on Intelligent Robots and Systems, 2009.

[20] M. Ani Hsieh, C., Anthony , V. Kumar, and J. Camillo. "Maintaining network connectivity and performance in robot teams," Research articles. Journal Field Robot, (2008).

[21] M. M. Zavlanos,. and G. J. Pappas. "Distributed connectivity control of mobile networks," in IEEE Transactions on Robotics, Dec. 2008, vol. 24, no. 6, pp. 1416 --1428.

[22] Ji. Meng, and M. Egerstedt.” Distributed coordination control of multiagent systems while preserving connectedness,” in IEEE Transactions on Robotics, Aug. 2007, vol. 23, no. 4, pp. 693 --703.

[23] M. A. Hsieh, A. Cowley, V. Kumar, and CJ. Taylor.” Towards the deployment of a mobile robot network with end-to-end performance guarantees,” in Proceedings IEEE International Conference on Robotics and Automation (ICRA), May 2006, pp. 2085 --2090.

[24] F. Zeiger, N. Kraemer, and K. Schilling. “Commanding mobile robots via ad-hoc wireless networks - a comparison of four ad-hoc routing protocol implementations,” in IEEE International Conference on Robotics and Automation (ICRA).

[25] Y. Mostofi, M. Malmirchegini, and A. Ghaffarkhah.” Estimation of communication signal strength in robotic networks,” in IEEE International Conference on Robotics and Automation (ICRA), May 2010, pp. 1946 --1951.

[26] D. P. Spanos and R. M. Murray. "Robust connectivity of networked vehicles"' in IEEE Conference on Decision and Control Conference, Dec. 2004, vol.~3, pp. 2893 -- 2898 Vol.3.

[27] M. Ayad, and, Mo. Ayad, " Investigates the Radio Frequency Obstacle Effects on Wireless Signal for Robot Communication, "in Future Technologies Conference, FTC 2021, Oct. 2021.

[28] M. Ayad, and R. Voyles, "R.F. Mapping for Sensor Nodes Connectivity and Communication Signal Recovery," Computing Conference 2021, London, United Kingdom, July 2021.

[29] CST2021, CST Studio Suite 3D EM simulation and analysis software.

[30] M. Ayad, and R. Voyles, "Message Routing and Link Maintenance Integration for Robotic Network Connectivity," in Future Technology Conference (FTC- 2020), Vancouver, Canada, Nov. 2020.

[31] M. Ayad, J. J. Zhang, R. Voyles, M. H. Mahoor. "Mobile robot connectivity maintenance based on RF mapping," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2013.

[32] D.Jurafsky. and J. H. Martin. Speech and language processing,,2000, in Prentice Hall, Inc.

[33] Maya Nayak and Bhawani Sankar Panigrahi ." Advanced Signal Processing Techniques for Feature Extraction in Data Mining," in International Journal of Computer Applications, April 2011, vol. 19, no. 9, pp. 30-37.

[34] M. Ayd, J. J. Zhang, R. Voyles and M. H. Mahoor.” Electromagnetic Field Recognition for Proactive Robot Communication Connectivity Maintenance,” 2012 Asilomar Conference on Signal, Systems, and Computers (Asilomar), 2012.

[35] M. A. Ayad,” Meta-Routing: Synergistic Merging of Message Routing and Link maintenance,” at https://digitalcommons.du. edu/etd/40.

A Review of Mobility Supporting Tunneling Protocols in Wireless Cellular Networks

Zeeshan Abbas, Wonyong Yoon
Department of Electronics Engineering
Dong-A University, Busan
South Korea

***Abstract*—With recent technology advancements mobility support is one of the major needed parameters by any wireless or mobile networks. Continuous mobile movement from one cell to another or from one network to another requires continuous mobility support. Previously, tunneling protocols employment was the technique to support UE's inter or intra network mobility. More specifically, GRE, GTP, MIPv6 or PMIPv6 were employed for mobility support. In tunneling encapsulation of one protocol over another protocol is performed to deliver IP packet during inter network or intra network handover. In terms of usage scenario of each tunneling protocol, tunnel establishment, data transfer and tunnel release, an overview and comparison of tunneling protocols is presented in this paper. 3GPP and WLAN interworking, and GAN based usage scenarios and supported tunneling mechanisms has been discussed. Some insights regarding security, multiplexing, multiprotocol and packet sequencing support are also provided for each tunneling protocol.**

***Keywords*—Tunneling; mobility; 3GPP; WLAN; interworking**

I. Introduction

With the recent advancements in wireless and mobile networks there is need of mechanisms that can support coexistence of multiple radio access technologies. These technologies should not also coexist but also provide seamless mobility between different radio access technologies. Looking on this thing various developments has been performed from different researchers. However, basically mechanism that provides support seamless handover are depends on tunnel that is created between different radio access technologies for establishing connectivity to the core network. Generic Routing Encapsulation (GRE) is one of the pioneering tunneling protocols that provide support for switching from one radio access technology to another. This protocol is based on encapsulation of one protocol over another protocol. Proxy Mobile IPv6 (PMIPv6) is mostly used for encapsulation. It is also specified in standard as IP session continuity signaling being used by Evolved Packet Core (EPC). This protocol supports mobility of terminals or UEs for various radio access technologies, e.g., Long Term Evolution (LTE) radio access, WLAN, 3G radio access, WiMAX and radio standards from 3GPP2. Basically, protocols just like PMIPv6 and MIPv6 manage the path for IP packets destined for different network or radio access technology. This kind of mobility approaches not also support seamless handover but also ensures efficient utilization of network resources, user privacy and network security.

Another type of tunneling protocol being used for supporting terminals mobility is GPRS Tunneling protocol (GTP). This protocol is also being used between different 3GPP core network entities. In which once a tunnel is established between different network entities then IP packets can be encapsulated and tunneled between these network entities. IP Security (IPSec) is another tunneling protocol used for protecting data integrity of wireless devices being delivered to the core network entities. This is the security association mechanism in which mutual authentication between terminal and access gateway is performed. Negotiation of security keys for a connection is also performed. Packets in IPSec are encrypted and encapsulated within a new packet with new control information and are delivered by IPSec tunnel from terminal or user equipment to access gateway.

II. Background

In Fig. 1, we have tried to cover architectural elements needed for supporting different type of technologies. As we can see in upper part of the figure 2G network components that will be needed for Voice Call Continuity (VCC) which provides support for anchoring circuit switched voice call in IP infrastructure by transferring speech path between these two domains transparently to end user [1, 2]. Tunnel is needed when VCC supported UE moves from 3G/4G network to 2G network then signaling messages between MME and Interworking solution Function for 3GPP2 1xCS (1xCS IWS) are transferred using S102 Tunnel. Another tunnel named IPSec is being used in Generic Access Network (GAN) and Wireless Local Area Network (WLAN) between UE to GAN Controller and ePDG network components respectively. GTP, PMIPv6 and GRE tunnels are used for secure data transfer between different network components like SGW, PGW, SGSN MME from 3G or 4G networking technologies, respectively.

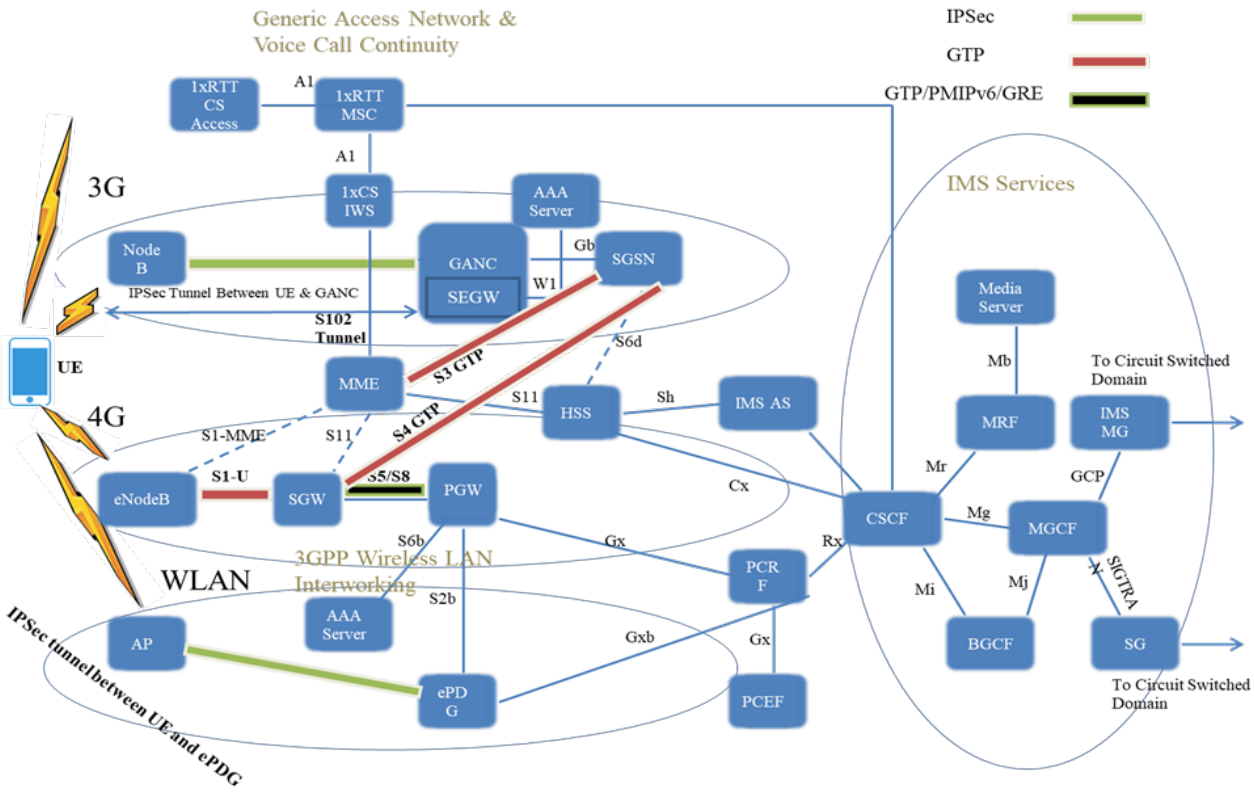


Fig. 1. Generic Heterogeneous Networks and Interworking Architecture.

However, authors in [3, 4] surveyed various tunneling protocols for supporting mobility. These tunneling protocols are IP based tunneling protocols used to support mobility in IPv6 based networks. Implementation of these protocols depends on usage scenario, as scenario can be host based, network based, soft handoff or hard handoff based Micro and Macro Mobility usage scenarios. Depending on these each different protocol is discussed for supporting seamless handover and mobility support in IPv6 networks [5]. However, some authors discussed these protocols from centralized or distributed point of view. Implementation of these protocols can be dependent on centralized or distributed support of these protocols. Authors suggested that use of protocols like GTP and PMIPv6 in centralized fashion may be not an effective solution because of single point of failure and scalability issues. So, they think that some distributed usage scenarios should be discussed. However, they mentioned the few by employing some already defined techniques just like de coupling control and user plane.

III. Tunneling Protocols for 3GPP

A. Generic Routing Encapsulation (GRE)

Currently, various protocols support encapsulation of one protocol over another. Generic Routing encapsulation (GRE) is among one of the encapsulation protocols which provide support for encapsulation of one protocol over another protocol. Simple IP packets are encapsulated in GRE header and transmitted to different intervening routers [6, 7].

1) Usage scenarios: GRE tunnel implementation is based on GRE encapsulation of data from network entity to another network entity over some other mobility supported protocol i.e., PMIPv6. Basically, GRE is used with PMIPv6 that can support mobility for UE if it is moved from one network to another. Then data can be transferred by using GRE encryption and PMIPv6 mobility support option [8].

2) Tunnel establishment and data transfer: Fig. 2 illustrates GRE tunnel establishment procedure. By adopting PMIPv6 as mobility support protocol and employing GRE as data encryption technique for security purposes a tunnel can be established between as Mobility Access Gateway (MAG) and Local Mobility Anchor (LMA). GRE key option is needed for establishing a tunnel between MAG and LMA. GRE keys are exchanged between these two entities by using Proxy Binding Update (PBU) defined in PMIPv6. MAG and LMA establish a GRE tunnel by the agreed GRE keys to transmit uplink and downlink traffic [9].

3) Tunnel release: Once a tunnel is established for data transfer it can also be removed from network. By following some steps tunnel between MAG and LMA can removed. MAG transmits a PBU message to LMA for release of the LMA binding. LMA completes release of binding by transmitting an acknowledgment message to MAG. During tunnel release process all resources just like IPv6 Home Network Prefix, IPv4 Home Address, Downlink and Uplink GRE Key, GRE Tunnel Tear-down deletion and de-assignment is performed [6].

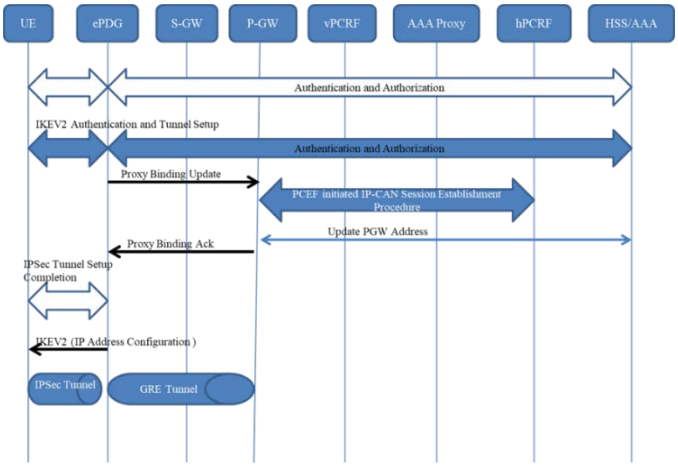


Fig. 2. An Initial Attachment over GRE and Tunnels Establishment.

B. GPRS Tunneling Protocol (GTP)

For transporting IP packets with in network or outside network GPRS tunnels concept is used. In which a tunnel is established between different network entities for successful transmission of IP packets within the network. For this purpose, a tunnel is established between different end points for data transmission and a unique identity named Tunnel End Point Identifier (TEID) is assigned to IP packets. Packets being received at different end points are forwarded based on their TEID's. According to 3GPP technical documents there are two types of protocols used in GTP one is GTP-U [6] and other is GTP-C [10]. GTP-U is used for delivering user data to different network entities. However, GTP-C is used to exchange control plane messages among different network entities.

1) Usage scenarios: GPRS Tunneling protocol implementation can be observed in various scenarios depending on type of interfaces being used. As, previously in [11] enhanced GTP was employed for GSM network which adopted Packet Data Protocol (PDP) to reduce tunneling overhead being faced at that time. Basically, one usage scenario that can be observed in different standard documents is 3GPP wireless LAN interworking. GTP can be implemented on different interfaces for providing connectivity between different network interfaces. GTP implementation using 3GPP plus trusted Non-3GPP access over S2a interface or 3GPP plus untrusted Non-3GPP access over S2b are provided to support usage scenarios. Seamless IP session continuity is supported between cellular networks and wireless local networks without the change of the address [12].

2) Tunnel establishment and data transfer: Fig. 3 illustrates GTP tunnel establishment procedure. First of all, Non-3GPP IP Access specific procedures takes place after that UE Authentication and authorization is performed by looking into HSS where subscription information of the user/subscriber is stored. 3GPP AAA transmits a reply to trusted non-3GPP network with information on all the authorized APNs and additional PDN GW selection. Upon completion of authentication and authorization, L3 attach procedure specific to non-3GPP access is initiated. UE transmits a request for session start from the obtained list of available APNs. Otherwise, PDN gateway selection procedure takes place. If UE connection to the particular APN becomes successful, trusted non-3GPP network transmits a message for making a connection (IMSI, APN, RAT type, Trusted non-3GPP IP Access TEID, etc.) message to PGW. After that PDN GW initiates the IP-CAN Session Establishment Procedure with the PCRF. PCRF creates IP-CAN session related information and responds to the PDN GW with PCC rules and event triggers. Then, the selected PDN GW informs 3GPP AAA server about PDN GW identity and the APN corresponding to the UE's PDN connection and also information about selected S2a protocol (GTP). PDN GW replies with a create session response (PDN GW Address for the user plane, PDN GW TEID of the user plane, PDN GW TEID of the control plane, PDN Type, PDN Address, EPS Bearer Identity, EPS Bearer QoS, APN-AMBR, Protocol Configuration Options and Cause message to the Trusted non-3GPP IP Access, with IP address assigned to the UE. Then, GTP tunnel is set up between trusted Non-3GPP network and the packet gateway. Once a tunnel is established, data transfer can take place [12].

3) Tunnel release: To release the tunnel or detach procedure is accomplished by following a procedure. First of all, a mobile or Trusted Non-3GPP network starts an access specific detach procedure from the network. Access technology specific detach trigger procedure is performed for tunnel release. Then, active bearers for UE and PDN connection are deactivated by the trusted non-3GPP IP Access sending a Delete Session Request to the gateway. Then the gateway informs AAA of PDN detach. PDN Gateway on receiving message deletes IP session associated with that particular UE. PDN Gateway and PCRF perform PCEF-Initiated IP CAN Session Termination Procedure. Then PDN Gateway sends a message to acknowledge with delete session response message. Finally, resources for trusted Non-3GPP network will be freed [9].

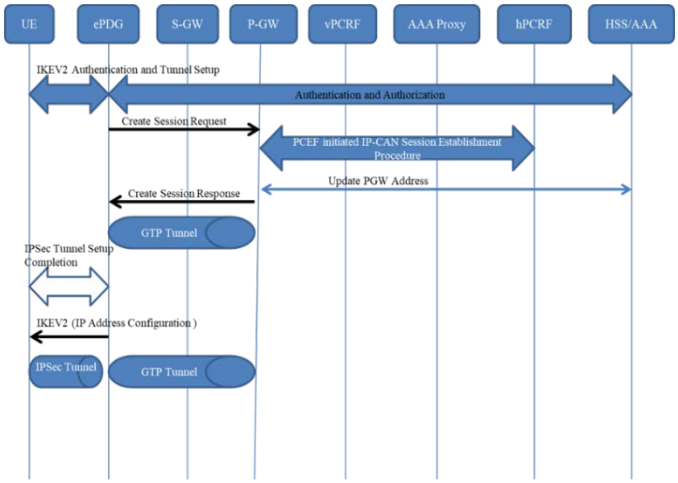


Fig. 3. An Initial Attachment over GTP and Tunnels Establishment.

C. Proxy Mobile IPv6

For supporting mobility of end users in different networks a protocol was introduced named Mobile IPv6. This protocol allows the users to keep online with mobility within IPv6 network. The basic concept behind this protocol is each node has a home address when connected to some network. However, by moving to another network it is associated to other network based on Care-of–Address (CoA). The address provides information on UE's latest point-of-contact. The protocol supports IPv6 nodes to store information on UE's home address and CoA. It also allows to send packets destined for UE by utilizing UE CoA. Tunneling mechanisms like GTP or GRE can also be used to support MIPV6 data transfer [13].

In network-initiated mobility control, network side maintains of the location of UE and triggers the necessary mobility message exchange. A proxy mobility agent is responsible for performing the mobility signaling with home agent. UE upon reaching to another network will try to connect to an access link. MAG over the link will perform authorization for network-based mobility. After authorization UE can perform address configuration and can move anywhere in PMIPv6 domain [14]. Authors in [15] presented some analytical studies regarding PMIPv6 where LMA is placed far from current MAG. In that case PMIPv6 will suffer from significant handover delay. So, based on mechanisms for supporting mobility whether predictive or reactive the authors in [15-18] suggested some enhancements to reduce handover latency, signaling cost and network utilization. Similarly, authors in [19-22] investigated messaging, data transmission, and tunneling overhead of different protocols with respect to PMIPv6.

1) Usage scenarios: Proxy Mobile IPv6 implementation is basically for supporting for transfer of control information between different network entities whether it is 3GPP or WLAN for handling network-based mobility. Basically, GRE is used with PMIPv6 that can support mobility for UE if it is moved from one network to another. Then data can be transferred by using GRE encryption and PMIPv6 mobility support option.

2) Tunnel establishment and data transfer: Fig. 4 illustrates PMIPv6 tunnel establishment procedure. For establishing tunnel between different network entities PMIPv6 follows some procedure in order to establish connection. Entities deployed with PMIPv6 protocol are named as MAG and LMA. MAG entity acts as SGW in 3GPP or ePDG in WLAN network environment. MAG initiates the tunnel establishment procedure for UE attach for the first time. MAG first transmits a PBU with APN to LMA. This results in LMA binding for UE's PDN attach. LMA completes binding by transmitting an acknowledgment to MAG. In the case of multiple PDN support for a single APN, each PDN connection ID is included in the acknowledgment. MAG generates a downlink GRE key distinct from any existing connection. MAG assigns a Fully Qualified PDN Connection Set Identifier. This identifies a group of PDN connections belonging to a group of UEs. MAG Includes LMA User Plane Address Mobility Option if the MAG supports the capability to receive from the LMA an alternate LMA address for user plan. On PBU reception, LMA selects the PDN based on APN information delivered in the PBU. LMA allocates an IPv4 or IPv6 address on receipt of PBU. Also, LMA generates uplink GRE key distinct from existing PDN connection's uplink traffic for that UE. After tunnel is established between different network entities data can be transferred by using any type of encapsulation, i.e. GRE [6].

3) Tunnel release: MAG initiates the release of PDN connection to tear down an existing PDN connection with LMA. MAG first transmits a PBU to LMA to delete the LMA binding for the UE's PDN connection. LMA completes deletion of the binding by transmitting a PBA to MAG. During tunnel release procedure, all resources such as IPv6 prefix, IPv4 address, Downlink and Uplink GRE Key, GRE Tunnel Tear-down deletion and de-assignment is performed [6].

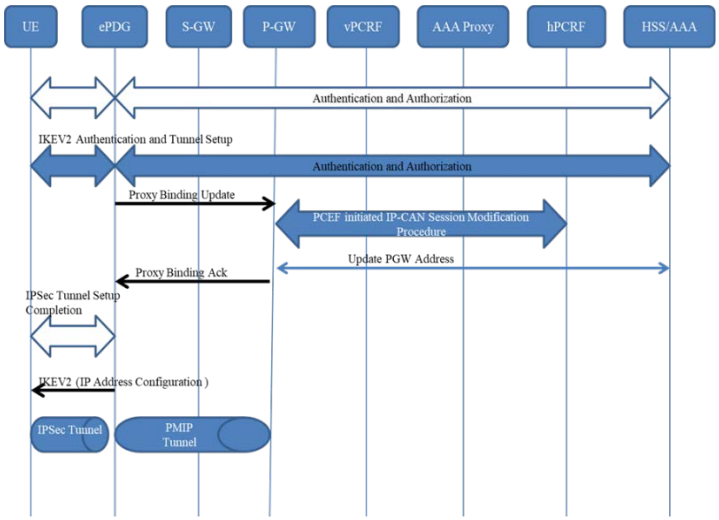


Fig. 4. An Initial Attachment over PMIPv6 and Tunnels Establishment.

D. Dual Stack Mobile IPv6

Dual Stack Mobile IPv6 (DSMIPv6) is an extension to MIPv6 functions. It allows UEs to request their home agent to forward IP packets addressed to their home address, and to their IP care-of address. A dual stack mobile can simultaneously enable both IPv4 and IPv6. Hence, two different mobility protocols are not needed at the same time [23, 24].

1) Usage scenarios: Usage scenarios can be based on for either 3GPP-WLAN interworking or mobility support within 3GPP. Solutions that enable seamless mobility between 3GPP-WLAN interworking and within 3GGP are needed such that current 3GPP based packet sessions can be served without interruption to UE's received service perception during the change of access network [25].

2) Tunnel establishment and data transfer: If UE is already on 3GPP Access and discovers the 3GPP I-WLAN domain, it may decide to change point-of-contact to

discovered 3GPP I-WLAN. For that purpose, UE will establish an IPSec Tunnel with ePDG. After establishing IPSec tunnel UE sends Binding Update message to HA and informs HA of its IP address. As a result of BU, DSMIPv6 tunnel is set up between UE and HA. After tunnel establishment, data plane messages can be transmitted using I-WLAN [25]. Fig. 5 illustrates DSMIPv6 tunnel establishment procedure.

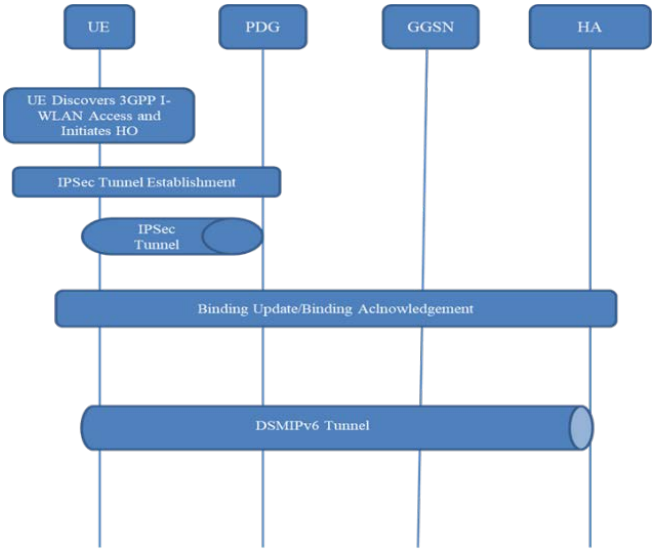


Fig. 5. An Initial Attachment over DSMIPv6 and Tunnels Establishment.

E. IP Secure Tunnel (IPSec)

For securing data integrity and for transmission another form encryption protocol is IPSec. This protocol provides support to encapsulate original IP packet and assign a new IP header to deliver the encapsulated data to other side of the network. Authentication header is also included in together with ESP when IPSec is being used in tunnel mode [26-28] for providing security and mobility support.

1) Usage scenarios: UE with connectivity to trusted or un-trusted non-3GPP access needs some security architecture to connect to 3GPP EPS. User identity confidentiality and devices identity confidentiality is needed in 3GPP EPS and Non-3GPP access for providing connectivity for non-3GPP access devices to 3GPP EPS [29].

2) Tunnel establishment and data transfer: If connection of UE is already established with non-3GPP access network then that UE can also access 3GPP access network by using this security mechanism, during which a secure tunnel is established. An example of such tunnel establishment is one between UE and ePDG. They first exchange some messages known as IKE_SA_INIT. Then UE sends user identity and APN information in IKE_AUTH step. It initiates negotiation of child security associations. ePDG sends Authentication and Authorization Request message to the 3GPP AAA Server containing user identity and APN information. Then 3GPP AAA server checks authentication vectors from HSS/HLR, and stores the information like IMSI and EAP-AKA requested authentication method in HSS. Then, 3GPP AAA server initiates authentication challenge by transmitting a reply to ePDG. ePDG then responds with its identity and a certificate. It sends the AUTH parameter to protect the previous message it sent to the UE. Message sent by 3GPP AAA server is also attached in that response message. UE checks the authentication parameters and responds to the authentication challenge. Then, the ePDG transmits the EAP-Response/AKA-Challenge message to the 3GPP AAA. AAA checks whether the authentication response is correct. If everything is correct, AAA shall initiate the Subscriber Profile Retrieval. It registers to the HSS and checks user's subscription whether it is authorized for non-3GPP access. If all checks are done, AAA transmits a final answer to ePDG. Information consists of MSK (EAP-Master-Session-Key-AVP). ePDG uses the MSK to authenticate IKE_SA_INIT step signalings. The EAP Success/Failure message is delivered to the UE over IKEv2. UE shall take its own copy of the MSK as input to generate the AUTH parameter to authenticate the first IKE_SA_INIT message. ePDG checks the correctness of the AUTH sent by the UE. ePDG calculates the AUTH parameter which authenticates the second IKE_SA_INIT message. Then, AUTH parameter is sent to the UE together with the configuration payload, security associations and the rest of the IKEv2 parameters. Now IKEv2 negotiation ends [29].

In Table I given below, we have some insights of each tunneling technique. First of all is security mechanism that should be supported by each tunneling protocol. Of the above-mentioned protocols IPSec supports complete security mechanism. In IPSec, Authentication Header (AH) protocol provides data origin authentication. Encapsulating Security Payload (ESP) supports data confidentiality [26]. Similarly, GRE provides some security by providing a four-byte key field for the purpose of origin authentication [3]. While regarding GTP security can be provided by assigning a unique tunnel end point identifier. Another key feature for each tunneling protocol is support for multiplexing which means supporting multiple simultaneous end devices. Separate tunnels may be set up; however separate tunnels impose processing overhead and increased delay for tunnel establishment. So, a better option is to share one tunnel among all end devices. A unique field is needed in tunneling IPSec provides this by Security parameter index. However, GRE and GTP provide multiplexing support by using GRE key field and GTP TEID field, respectively [6, 12]. Multiprotocol support is also needed by each tunneling protocol. GRE provides multiprotocol support as it was defined as general encapsulation protocol. However, IPSec fails to provide support for multiple protocols. Another important parameter for each protocol is packet sequencing. IPSec has sequence number such that in-order delivery of packets can be feasible.

TABLE I. COMPARISON OF TUNNELING PROTOCOLS

Protocol	Security	Tunnel Establishment and Configuration	Tunnel Management	Support for Multiplexing	Support for Multi-protocol	Support for Packets Sequence	Standardizing Body	Overview
GRE [6, 30, 31]	Yes (four-byte key field)	Network Explicit	No	Yes (Using Key field)	Yes	No	IETF	Encapsulation of one protocol over another protocol
GTP [12, 32]	Yes (IPSec ESP with encryption and integrity protection)	Network management Explicit	Yes (Create or Delete Session or bearer)	Yes (TEID will provide that kind of support)	Yes	Yes	3GPP	IP based Transport Protocol
PMIPv6 [14, 33]	Yes (chained-tunnel approach provides hop-by-hop based security protection)	Network Management Explicit (PBU & PBA)	Yes (PBU and PBA containing Uplink and Downlink GRE keys stored in Binding Cache)	Yes (GRE key option will provide support)	Yes (PIMPv6 with GRE encapsulation)	Yes (GRE key field)	IETF	Localized mobility management
DSMIPv6 [23, 33]	Yes (IKEv2 based IPSec Security Association)	Client or Host initiated tunnel establishment (PBU & PBA)	Yes (PBU containing Key Management Mobility Option)	No (Attach needs to be done for separate PDNs)	Yes (I-WLAN Attach with mobility service)	No	IETF	Support MN roaming over IPV6 or IPv4 networks and transmission of IPv4/v6 packets over the tunnel to HA
IPSec [26, 30, 31]	Yes (complete build in security)	IKE interchange Implicit	No	Yes (via Security Parameter Index)	No	Yes (sequence number field)	IETF	Security and protecting data integrity

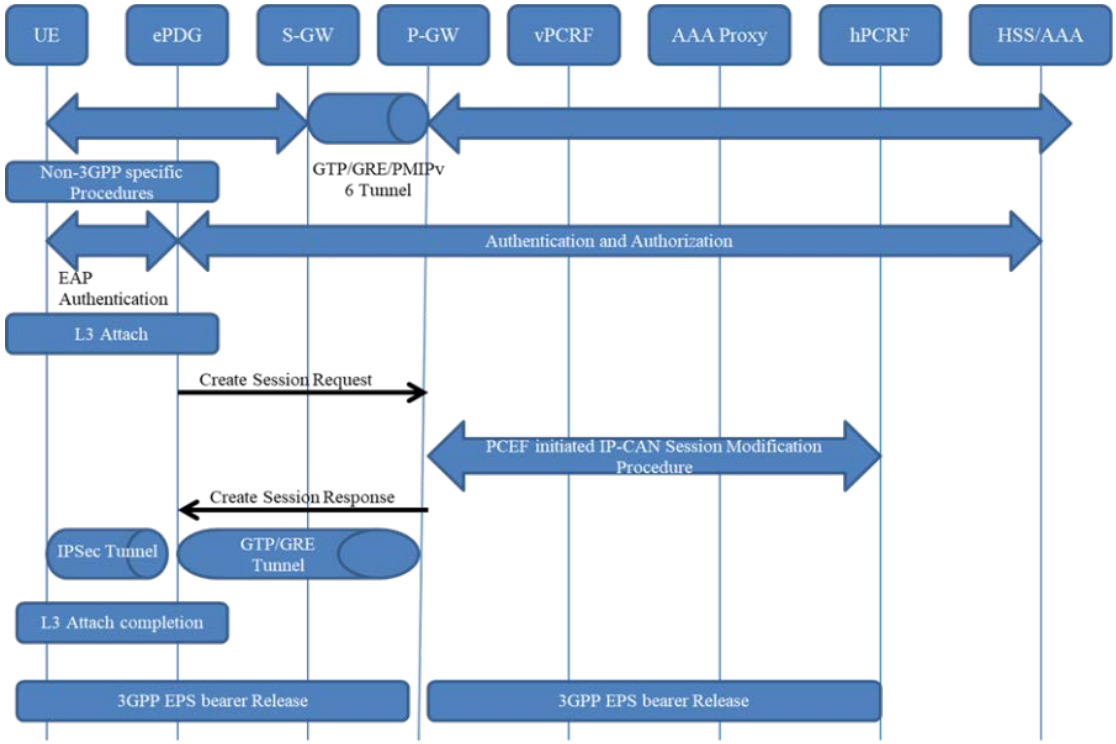


Fig. 6. A Simple 3GPP to WLAN Session Mobility Scenario and Tunnels Establishment.

Fig. 6 represents a generic session mobility scenario in which different tunneling protocols can be employed for handling mobility and data transfer. Initially Non-3GPP intrinsic L2 signaling is completed. Then EAP authentication procedure is started between UE Trusted Non-3GPP and AAA. As authentication reply by AAA, a group of all the authorized APNs plus additional PDN gateway selection is returned to the access gateway. Upon completion of authentication and authorization, non-3GPP radio specific L3 attach procedure is

initiated. Then, ePDG transmits a message to establish a data session to the gateway. PDN gateway initiates the IP CAN Session Establishment Procedure with the PCRF and the PCRF provides the APN-AMBR and Default Bearer QoS to the PDN GW in the response message. In response to the create session request message PGW sends a create session response message caching the all required information. After that session is established successfully and ePDG is also authenticated by the UE. IP session between UE and P-GW is now established. Packets from UE to ePDG are tunneled using IPSec Tunnel and then onward using some other tunneling protocol i.e., GRE or GTP [34].

IV. Tunneling Protocols for 3GPP-WLAN Interworking

Radius or diameter based 3GPP WLAN interworking is discussed in [35] which represents some network some network components WLAN UE, WLAN Access Network (WLAN AN), 3GPP AAA Server and Home Subscriber Server (HSS) database. To connect to WLAN, WLAN UE uses a SIM or USIM (UMTS Subscriber Identity Module) containing the authentication keys for the mobile subscriber. To connect UE to 3GPP network, WLAN Access Network (WLAN AN) plays a role of anchor between them. 3GPP AAA performs authentication and authorization for UE. When the WLAN AN receives a WLAN UE connection request, it may perform an initial access negotiation with the WLAN UE to obtain identity information and then pass this information to the 3GPP AAA server as part of an authentication/authorization request. WLAN AN may be RADIUS or Diameter-based. The 3GPP AAA Server matches data from the authentication/ authorization request with information in a trusted database, called a Home Subscriber Server (HSS). If a match is found, and the subscriber's credentials are correct, the 3GPP AAA server responds with a reply to WLAN AN. This indicates the acceptance of the request. Otherwise or if a problem is found with the subscriber's credentials, 3GPP AAA returns a reject message. This results in the termination of WLAN UE connection.

Upon establishment of WLAN UE connection, WLAN AN may forward accounting information to 3GPP AAA to record the transaction for future price charging. 3GPP AAA must be able to translate between RADIUS and Diameter (and vice versa), to support connection requests from legacy RADIUS WLAN ANs to the 3GPP network which uses the Diameter protocol. HSS stores subscriber data like keys to complete authentication to allow access for user device. It also performs authorization to enable access to limited service and functions. Overall for interworking support tunneling is the mechanism adopted as discussed in [36]. Also, another mechanism that supports offloading end users from Radio Access Network to WLAN using IPSec transport mode has been discussed in [37] to reduce networking security overhead caused by IPSec Tunnel mode. Fig. 6 illustrates 3GPP to WLAN mobility procedure and required tunnel establishment.

A. Directly Accessing to the Internet

In this case, the Internet is connected through WLAN AN. Users access WLAN AN. IMS AAA is responsible for authentication of users. It uses either the EAP-SIM or EAP-AKA protocol that originates from RADIUS and Diameter WLAN ANs. The SIM-authentication mechanism is used against the subscriber information stored in the HSS. Authentication is performed directly from the WLAN AN. Upon completion of authentication, authorization will return policy information for session establishment [35, 38].

B. Accessing through 3GPP

In this case, users can access connection service to the Internet via a secure tunnel to 3GPP IMS. IP packets are forwarded through tunnel to 3GPP IMS network via WLAN Access Gateway (WAG) and ePDG. WAG acts as a dynamically configured firewall. ePDG is a tunnel end-point. Multiple tunnels are possible to support any number of simultaneous services. ePDG requests authorization separately from the authentication request. For example, WLAN UE may initiate a tunnel towards the ePDG. This is followed by authentication and tunnel establishment [35].

C. Generic Access Network (GAN)

GAN was developed as an advancement of Unlicensed Mobile Access. GAN is another type of network that can coexist with 3GPP core network. User equipment with multiple radios (WLAN and 3GPP) can access 3GPP through GAN. GAN is applied usually through IEEE 802.11 WLAN. Gateway in 3GPP core named as Generic Access Network Controller (GANC) is responsible for handling traffic coming from WLAN. Initially, UE starts working on by default 3G settings when powered on using WLAN. UE connects to appropriate AP. IP address is configured to perform GAN discovery. It establishes an IPSec tunnel with Security Gateway (SeGW). It registers with GANC. If GANC accepts connection UE GAN mode is enabled [39, 40].

1) GAN Discovery and registration: First of all, for GAN mode selection is done during Discovery phase. MS transfers its GAN Mode Support information to GANC. GANC can assign appropriate port on default GANC based on the GAN mode support information provided by MS. During discovery phase, MS obtains the address of default GANC. It also discovers that of associated SEGW. Then it establishes a secure IPSec Tunnel. After sending IP address query for both SEGW and GANC, MS opens a TCP session with GANC. GANC responses with Discovery accept message. After discovery phase, MS initiates registration with default GANC, which can act as the serving GANC after establishment of connection and registration procedure. GAN registration procedure confirms adequate registration of a mobile to the controller. The procedure helps the MS for appropriate GAN mode selection. After establishing a secure tunnel with SEGW and IP address has also been obtained. MS can then send a registration request message to GANC. Information contained in registration request message is current camped cell of MS i.e., GERAN/UTRAN/EUTRAN, Last LAI or TAI, IMSI and information about required GAN services. If GANC accepts registration request it responds by sending register accept message to MS.

Fig. 7 illustrates GAN tunnel establishment procedure using EAP/SIM(AKA) over IKE. In GAN different tunneling protocols are being supported on different levels of network. During establishment of connection to GANC UE establishes an IPSec tunnel with SEGW. The security association of IPSec tunnel is established. Another tunnel protocol used between GANC and GGSN is GPRS Tunneling protocol. This tunnel is established during data transfer from between UE and the network i.e., 2G or 3G. GA-RRC Packet Transport Channel is made for packet switching domain on both sides of network entities. Connection status of GA-RRC is active or inactive. Some triggers are used for GA-RRC PTC state activation. First one is when GANC receives RAB assignment message from GGSN and second one is when GANC receives relocation request from SGSN. When these two triggers happen SGSN includes the information like RAB ID, IP Address and GTP-U TEID in RAB Assignment Request or Relocation Request message sent to GANC. During that time GA-RRC channel on UE is activated by GANC by forwarding the received message from GGSN to UE. After channel activation at UE now that particular UE will be in GA-RRC-Connected PTC-ACTIVE sub state. RAB Assignment is transmitted to GGSN using the same information received already. Upon establishment of PDP context, a mobile may start transmission upward data in GA-RRC PDU. GANC relays the payload part of PDU to SGSN in Iu-PS G-PDU message. SGSN transmits downward user data in Iu-PS G-PDU toward GANC. The message includes MS TEID already received during RAB Assignment Request or Relocation Request messages [39]. On the other hand, previously some work also focused on mobility management. Mechanism for supporting GAN handoff is presented in [41], in which authors focused on adaptive keep alive interval messages for allocation of resources and mobility management and reducing the handoff failure probability. Similarly, authors in [42] presented a VoLGA based solution. They suggested to with some software and interface addition connectivity for VoLGA can be provided.

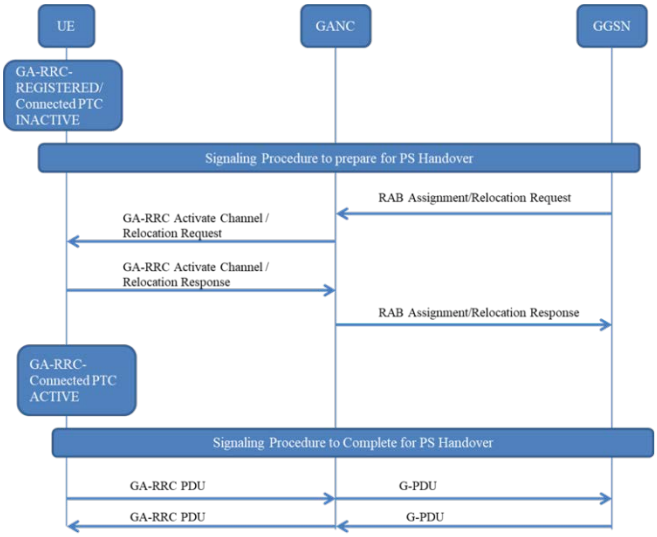


Fig. 7. Simple GAN RAB Assignment and Relocation Mechanism and Tunnel Establishment.

V. Conclusion

In this paper, an in-depth review of various tunneling protocols employed by various wireless and mobile networks for supporting mobility is presented. Tunnel establishment, data transfer, tunnel release and respective usage scenarios for each tunneling protocol has been considered for systematic and thorough comparison of existing tunneling protocols. An insight of each tunneling protocol is provided and discussed. Specifically, mechanisms like security, multiplexing support, multiprotocol support and packets sequencing support has been discussed. Focusing usage scenarios, tunneling mechanism being used in 3GPP wireless LAN interworking, and Generic Access, Network (GAN) has been discussed.

Acknowledgment

This work was supported by the Dong-A University research fund. W. Yoon is the corresponding author.

References

[1] A. K. Salkintzis, M. Hammer, I. Tanaka, and C. Wong, "Voice call handover mechanisms in next-generation 3GPP systems," Communications Magazine, IEEE, vol. 47, pp. 46-56, 2009.

[2] J. Namakoye and R. Van Olst, "Performance evaluation of a voice call handover scheme between LTE and UMTS," in AFRICON, 2011, 2011, pp. 1-5.

[3] D. Saha, A. Mukherjee, I. S. Misra, and M. Chakraborty, "Mobility support in IP: a survey of related protocols," Network, IEEE, vol. 18, pp. 34-40, 2004.

[4] W. Lili, G. Jianfeng, Y. Ilsun, Z. Huachun, G. Deyun, Y. Kangbin, and K. Pankoo, "Survey on distributed mobility management schemes for Proxy mobile IPv6," in Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th, 2014, pp. 132-138.

[5] N. Chuangchunsong, S. Kamolphiwong, T. Kamolphiwong, R. Elz, and P. Pongpaibool, "Performance evaluation of IPv4/IPv6 transition mechanisms: IPv4-in-IPv6 tunneling techniques," in Information Networking (ICOIN), 2014 International Conference on, 2014, pp. 238-243.

[6] D. Farinacci, T. Li, S. Hanks, D. Meyer, and a. P. Traina, "RFC 2784, Generic Routing Encapsulation (GRE)," RFC 2784, March 2000.

[7] G. Yu, L. Breslau, N. Duffield, and S. Sen, "GRE Encapsulated Multicast Probing: A Scalable Technique for Measuring One-Way Loss," in INFOCOM 2008. The 27th Conference on Computer Communications. IEEE, 2008, p. 1.

[8] 3GPP, "Proxy Mobile IPv6 (PMIPv6) based Mobility and Tunneling protocols," in 3GPP TS 29.275, ed., July 2020.

[9] 3GPP, "General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U)," in 3GPP TS 29.281, ed, September 2021.

[10] 3GPP, "Evolved General Packet Radio Service (GPRS) Tunneling Protocol for Control plane (GTPv2-C)," in 3GPP TS 29.274, ed, December 2021.

[11] T. Shiao-Li Charles, "Enhanced GTP: an efficient packet tunneling protocol for General Packet Radio Service," in Communications, 2001. ICC 2001. IEEE International Conference on, 2001, pp. 2819-2823 vol.9.

[12] 3GPP, "Study on S2a Mobility based on GPRS Tunnelling Protocol (GTP) and Wireless Local Area Network (WLAN) access to the Enhanced Packet Core (EPC) network (SaMOG)," in 3GPP TS 23.852, ed, September 2013.

[13] C. Perkins, Ed., Johnson, D., and J. Arkko, "RFC 6275 Mobility Support in IPv6," RFC 6275, July 2011.

[14] S. Gundavelli, Ed., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "RFC 5213,Proxy Mobile IPv6",," RFC 5213, August 2008.

[15] L. Jun and F. Xiaoming, "Evaluating the Benefits of Introducing PMIPv6 for Localized Mobility Management," in Wireless

Communications and Mobile Computing Conference, 2008. IWCMC '08. International, 2008, pp. 74-80.

[16] J. Inwhee and L. Hyojin, "An efficient inter-domain handover scheme with minimized latency for PMIPv6," in Computing, Networking and Communications (ICNC), 2012 International Conference on, 2012, pp. 332-336.

[17] L. Meng-Hsuan, C. Whai-En, and H. Chao-Hsi, "HF-PMIPv6: An enhanced fast handovers for network-based mobility management," in Advanced Infocom Technology 2011 (ICAIT 2011), International Conference on, 2011, pp. 1-7.

[18] A. Rasem, M. St-Hilaire, and C. Makaya, "A comparative analysis of predictive and reactive mode of optimized PMIPv6," in Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International, 2012, pp. 722-727.

[19] H. Ali-Ahmad, M. Ouzzif, P. Bertin, and X. Lagrange, "Comparative performance analysis on dynamic mobility anchoring and proxy mobile IPv6," in Wireless Personal Multimedia Communications (WPMC), 2012 15th International Symposium on, 2012, pp. 653-657.

[20] P. Seung Yoon and J. Jongpil, "On Pointer Forwarding Based Mobility Management for Cost-Optimized Proxy Mobile IPv6 Networks," in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013 Seventh International Conference on, 2013, pp. 29-36.

[21] G. Song, X. Wang, X. Li, J. Huo, and Y. Liu, "Cost Analysis of a Novel Mobility Management: Interworking between PMIPv6 and MIPv6," in Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on, 2012, pp. 1-4.

[22] L. Jong-Hyouk, Y. Zhiwei, J. M. Bonnin, and X. Lagrange, "Dynamic tunneling for network-based distributed mobility management coexisting with PMIPv6," in Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on, 2013, pp. 2995-3000.

[23] H. Soliman, Ed., "Mobile IPv6 Support for Dual Stack Hosts and Routers", RFC 5555, June 2009.

[24] K. Mitsuya, R. Wakikawa, and J. Murai, "Implementation and Evaluation of Dual Stack Mobile IPv6," in Asia BSD Conference (AsiaBSDCon2007), 2007.

[25] 3GPP, "Mobility between 3GPP-Wireless Local Area Network (WLAN) interworking and 3GPP systems," in 3GPP TS 23.327 ed, September 2016.

[26] S. a. K. S. Kent, "RFC 4301, Security Architecture for the Internet Protocol," RFC 4301, December 2005.

[27] K. Byoung-Jo and S. Srinivasan, "Simple mobility support for IPsec tunnel mode," in Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th, 2003, pp. 1999-2003 Vol.3.

[28] J. Younchan and M. Peradilla, "Tunnel gateway satisfying mobility and security requirements of mobile and IP-based networks," Communications and Networks, Journal of, vol. 13, pp. 583-590, 2011.

[29] 3GPP, "3GPP System Architecture Evolution (SAE); Security Aspects of non-3GPP Accesses," in 3GPP TS 33.403, ed, July 2020.

[30] T. Saad, B. Alawieh, S. Guider, and H. T. Mouftah, "Tunneling techniques for end-to-end VPNs: generic deployment in an optical testbed environment," in Broadband Networks, 2005. BroadNets 2005. 2nd International Conference on, 2005, pp. 859-865 Vol. 2.

[31] A. Zhao, Y. Yuan, Y. Ji, and G. Gu, "Research on tunneling techniques in virtual private networks," in Communication Technology Proceedings, 2000. WCC - ICCT 2000. International Conference on, 2000, pp. 691-697 vol.1.

[32] 3GPP, "3G security; Network Domain Security (NDS); IP network layer security," in 3GPP TS 33.210, ed, July 2020.

[33] 3GPP, "3GPP System Architecture Evolution (SAE); Security aspects of non-3GPP accesses," in 3GPP TS 33.402 ed, July 2020.

[34] M. Crosnier, F. Planchou, R. Dhaou, and A. Beylot, "Handover Management Optimization for LTE Terrestrial Network with Satellite Backhaul," in Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd, 2011, pp. 1-5.

[35] 3GPP, "3GPP System to Wireless Local Area Network (WLAN) interworking; system Description," in 3GPP TS 23.234, ed, March 2015.

[36] G. Chai-Hien, L. Yung-Chun, Y. Shun-Neng, and L. Yi-Bing, "A Seamless Multi-link Switch Solution for LTE and Wi-Fi Integrated Networks," in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013 Seventh International Conference on, 2013, pp. 19-23.

[37] D. Migault, D. Palomares, E. Herbert, Y. Wei, G. Ganne, G. Arfaoui, and M. Laurent, "E2E: An Optimized IPsec Architecture for Secure and Fast Offload," in Availability, Reliability and Security (ARES), 2012 Seventh International Conference on, 2012, pp. 365-374.

[38] D. Celentano, A. Fresa, M. Longo, and A. L. Robustelli, "Improved Authentication for IMS Registration in 3G/WLAN Interworking," in Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on, 2007, pp. 1-5.

[39] 3GPP, "3GPP. Generic Access Network (GAN); Stage 2. ," in 3GPP TS 43.318, ed, July 2020.

[40] J. Kellokoski, "Challenges of the always-best-connected enablers for user equipment in Evolved Packet System," in Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2012 4th International Congress on, 2012, pp. 174-180.

[41] C. Kai-Hsiu and C. Jyh-Cheng, "Handoff Failure Analysis of Adaptive Keep-Alive Interval (AKI) in 3GPP Generic Access Network (GAN)," Wireless Communications, IEEE Transactions on, vol. 10, pp. 4226-4237, 2011.

[42] O. Stepaniuk, "Voice over LTE via Generic Access (VoLGA) as a possible solution of mobile networks transformation," in Modern Problems of Radio Engineering, Telecommunications and Computer, 2010.

Extended Kalman Filter Sensor Fusion in Practice for Mobile Robot Localization

Alaa Aldeen Housein, Gao Xingyu*, Weiming Li, Yang Huang
School of Electrical and Mechanical Engineering, Guilin University of Electronic Technology, Guilin, China

***Abstract*—Self-driving vehicles and autonomously guided robots could be very beneficial to today's civilization. However, the mobile robot's position must be accurately known, which referred as the localization with the task of tracking the dynamic position, in order for the robot to be active and useful. This paper presents a robot localization method with a known starting location by a real-time reconstructed environment model that represented as an occupancy grid map. The extended Kalman filter (EKF) is formulated as a nonlinear model-based estimator for fuse Odometry and a LIDAR range finder sensor. Because the occupancy grid map for the area is provided, just the inaccuracies of the LIDAR range finder will be considered. The experimental results on the "turtlebot" robot using robot operating system (ROS) show a significant improvement in the pose of the robot using the Kalman filter compared with sample Odometry. This paper also establishes the framework for using a Kalman filter for state estimation, providing all relevant mathematical equations for differential drive robot, this technique can be used to a variety of mobile robots.**

***Keywords*—Autonomous navigation; kalman filter; self-driving vehicle; simultaneous localization and mapping; occupancy grid map; ROS**

I. Introduction

Recently, mobile robots have been used to perform specialized tasks in a number of industries, including services, rescue, military, disaster relief, unmanned defense vehicles, and so on. Localization of mobile robots is a hard problem that many academics are seeking to address via the development of innovative techniques. Researchers have added more sources in order to build a powerful localization approach [1].Odometry is one of the most important techniques to tackle the posture tracking problem, it uses encoder data to track the motion progress from a specified beginning position. This technique tracks motion from a known beginning position using encoder data, the encoded data is sent to the central processor, which uses a geometric equation to update the robot's position [2, 3]. Due to a variety of factors such as wheel slippage, ground roughness, and varying wheel diameters, this approach has accumulative errors. So, under severe conditions, solely utilizing Odometry for localization virtually never results in an accurate state, and it becomes more active when other sources of sensing are used. Stereo, LIDAR range finder, sonar, compass, gyro, and GPS are the most often utilized extra sensors [4]. Building a robot from the scratch is expensive and time consuming, so working in an environment that guarantees theoretical study and practical implementation will be quiet helpful. ROS is an open source meta operating system that provides hardware abstraction, control implementation of commonly used functionalities, tools and libraries for building, writing, and running code for simulated and real robots after installing necessary drivers. The significance of this research rests in the framework it provides for fusing several sensors with a Kalman filter for robot localization, with the experimental results emphasizing notable reduction of errors in robot position. The paper is structured as follow: in section III the motion model of two wheeled robot was derived. Section IV presents Kalman filter pose tracking design. Section V discusses Kalman filter implementation results, followed by conclusion in section VI.

II. Literature Review

Iraj Hassanzadeh and Mehdi Abedinpour implemented an augmented unscented and extended Kalman filter for position tracking using a differential drive mobile robot with encoder readings, assuming real measurements are available. The work showed an improvement in pose tracking using this technique with the unscented filter outperforming the extended one [5]. Jaeyong Park and Sukgyu Lee investigated a mobile robot SLAM (simultaneous localization and mapping) technique based on EKF extended Kalman filter, with an additional extended Kalman filter used to enhance robot heading accuracy, because the robot's kinematic model was unclear due to the rough surface, its heading was deviated as it drove across uneven terrain. They proposed a method for correcting uncertain robot postures utilizing an extra extended Kalman filter on a simulation-based test [6]. The Kalman filter was applied on a Pioneer 2DX mobile robot by Edouard Ivanjko, Mario Vasak, and Ivan Petrovic, combining data from wheel encoders and a sonar sensor. The experimental result indicates that the kalman filter reduces posture tracking errors when compared to using Odometry alone [7]. Yusuke Misono and Yoshitaka Goto have implemented outdoor SLAM using Kalman filter in an outdoor environment using electric wheelchair mobile robot provided by LIDAR range finder and GPS the experimental results reveal a significant improvement of self-localization vehicle estimate by the SLAM algorithm compared with dead reckoning [8]. On a laser range finder-equipped robot platform, Angga Rusdinar, and Sungshin Kim investigated simultaneous localization and mapping with a particle filter. The experiment's findings showed that the suggested particle filter can improve map building performance and mobile robot localization accuracy inside indoor buildings [9]. Mohammed Faisal, Mansour Alsulaiman, and Ramdane Hedjar Hassan developed a localization method to reduce error accumulation in the dead-reckoning approach, since dead-reckoning is reliant on encoder information, they use an additional sensing source, the proposed localization

*Corresponding Author.

system uses the extended Kalman filter in combination with infrared sensors to enhance the mobility robots' localization, by rectifying errors in the robot's position and address the issue of dead-reckoning, with the working area's walls serving as references (landmarks)[10]. Yi-Xiang Wang and Ching-Lung Chang investigated SLAM under the robot operating system (ROS) utilizing a laser range finder (LIDAR), an Inertial Measurement Unit (IMU), an odometer, and an Ultra-wideband (UWB), and fused all of the above sensors using Extended Kalman filter. Experiment findings demonstrate that the mobile robot's average error distance in the system is restricted to 10cm [11]. For city navigation, Zanwu Xia and Si Tang developed a new technique to improve the accuracy of high definition (HD) maps in order to improve the localization of self-driving cars. The research focused on extracting the factors that have a high impact on the global map, such as feature sufficiency, layout, local similarity, and map representation quality. The Kalman filter was used to combine data from LIDAR, IMU, and GNSS systems. The experimental results show a reduction in accumulative errors [12].

In this study, we focus on the problem of indoor robot localization, which involves determining the position of the robot x, y, and its orientation Θ. The idea of the Kalman filter is to reduce the errors in both the mechanical model of the robot and the sensor readings. Kalman Filter is designed to deal with linear systems, but most nontrivial systems are nonlinear. Therefore, a new modified technique called extended Kalman filter (EKF) has been developed. This paper aims to deal with the uncertainties of a mobile robot by fusing Odometery and LIDAR range finder in the dead-reckoning method. Three steps are encoded in the method proposed here; the prediction step depends on the motion equation of the robot platform. Data acquisition by measurement sensors in the workspace, which are used to correct the robot position calculated in the motion prediction step. Update step that corrects the sum of motion uncertainty and measurement uncertainty, Fig. 1.

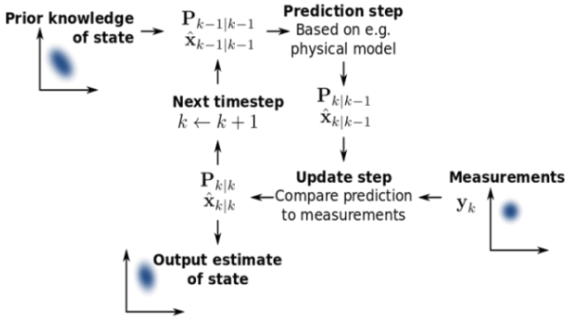


Fig. 1. Kalman Filter Estimator.

III. Motion Model

Regardless of whether robot position has to be adjusted or not, a kinematic model of robot motion exists that is dependent on the degree of freedom available to it. For example, a wheeled mobile robot without a manipulating arm (our robot platform) has three degrees of freedom (displacement along X-axis, displacement along Y-axis, and the orientation around Z-axis). Fly robot for instance has six degree of freedom (beside to displacement along orthogonal axis's X Y Z, there is orientation around each axis Θ_x Θ_y Θ_z).In the experiment, a two-wheeled mobile robot was used. Each wheel has encoders placed, besides the passive caster wheel for stability. The driving wheels are individually controlled. The following relationships define the mobile robot's kinematic model (Fig. 2).

$$x_{k+1} = x_k + d_k cos\Theta_k \quad (1)$$

$$y_{k+1} = y_k + d_k sin\Theta_k \quad (2)$$

$$\Theta_{k+1} = \Theta_k + \Delta\Theta_k \quad (3)$$

$$d_k = v_{t,k}.T \quad (4)$$

$$\Delta\Theta_k = w_k.T \quad (5)$$

$$v_{t,k} = \frac{v_{l,k}+v_{r,k}}{2} = \frac{w_{l,k}.R + w_{r,k}.R}{2} \quad (6)$$

$$w_k = \frac{v_{r,k}-v_{l,k}}{b} = \frac{w_{r,k}.R - w_{l,k}.R}{b} \quad (7)$$

Where x_k and y_k are the center gravity of robot platform ; d_k travel distance between two successive time interval k+1 and k ; $v_{t,k}$ mobile robot's translational velocity ; T is the sampling time ; Θ_k the robot's heading with the X-axis; $\Delta\Theta_k$ rotational angle of robot between k+1 and k time steps ; $v_{l,k}$ and $v_{r,k}$ the left and right wheel's respective linear velocities ; $w_{l,k}$ and $w_{r,k}$ the left and right wheel's respective angular velocities ; The two driving wheels have a radius of R ; b axle length or robot. The radius of both driving wheels, in sampling, is assumed to be equal. We add three more variables to (6) and (7) to account for sampling errors caused by not knowing the exact radius and axle length.

$$v_{t,k} = \frac{v_{l,k}+v_{r,k}}{2} = \frac{w_{l,k}.R + w_{r,k}.R}{2} \quad (8)$$

$$w_k = \frac{v_{r,k}-v_{l,k}}{b} = \frac{w_{r,k}.R - w_{l,k}.R}{b} \quad (9)$$

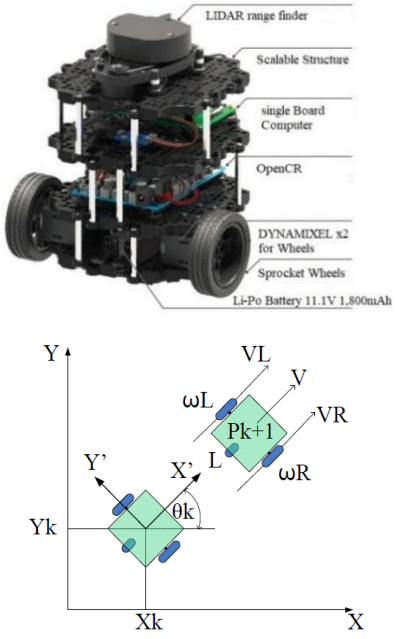


Fig. 2. Real Robot and its Kinematic Model.

The uncertainty of the exact wheel radius is represented by k_1 and k_2, while the uncertainty of the exact axle length is represented by k_3. In [13, 14, 15, 16] the systematic error correction approach is described in depth, as well as how the parameter values were calculated. Equations (1) to (7) describe the fundamental concept of Odometry position tracking.

IV. Location Tracking with Kalman Filter

Although there are many specific techniques for estimating the state of the system from, a set of measurements, most of these do not explicitly consider the noisy nature of measurements. This noise is typically described by statistics, which leads to have to use stochastic methods to tackle the problem[17, 18]. This section describes the stochastic state estimation process, initially, the basics of Bayesian filtering are presented, providing a brief mathematical derivation of how it is possible to make an estimate of the state. A mathematical treatment of the Kalman Algorithm for localization is then presented.

The difficulty in utilizing sensor fusion to localize a mobile robot is balancing the uncertainty of the state (x, y and Θ) with the LIDAR range measurement (robot's output) to achieve an optimum estimation of the posture. Kalman filter requires that the state random variables have Gaussian probability distributions that are adequately characterized by the mean and covariance [19, 20]. Time Update and Measurement Update are the two major stages in calculating the optimum state estimate. The state prediction is generated using the motion model based on the previous value and the control input value. The output forecasts are calculated using the measurement model based on the outcomes of the time update. The anticipated state mean and covariance are then adjusted by reducing the state covariance using the difference between

Expected and measured output.

A. Bayesian Filtering

A Bayesian filter is a mathematical tool that estimates the development of the system's state given the available data [21]. This tool necessitates:

The analytical knowledge of the transition function f_t and the stochastic knowledge of the noise of the state v_t

- The analytical knowledge of the output function h_t and the stochastic knowledge of the observation noise w_t
- The realization of the output of the system $z_{1:t}$ at time t.

Having this data, the Bayesian filter is able to estimate the function probability density. Once the state estimation problem has been formalized as a Baysian filtering problem, we have to find a mathematical formulation that allows us to return to the probability density function $p(x_{0:t}|z_{1:t})$ using our system knowledge and observations. First, we have to model the system's internal evolution as well as how it manifests itself through its observable outputs. Two conditional distributions are introduced for this purpose. The first is observation model $p(z_t|x_t)$, represents the density of the measurement z_t given the system state x_t .The second function $p(x_t|x_{t-1})$ is an evolution model that represents how the system develops over time. Using the Bayes rule and the Markovian chain assumption, it is possible to obtain an a posteriori probability density of the state incrementally. Because data arrives in stages over time, a recursive formulation of Bayes' rule known as the Bayes Filter is used. The a posteriori probability of the state is gradually refined in this formulation as measurements arrive.

$$p(x_{0:t+1}|z_{1:t+1})$$

$$= \frac{p(z_{t+1}|x_{0:t+1},z_{1:t})p(x_{0:t+1}|z_{1:t})}{p(z_{t+1}|z_{1:t})} \quad (10)$$

$$= \frac{p(z_{t+1}|x_{t+1})p(x_{0:t+1}|z_{1:t})}{p(z_{t+1}|z_{1:t})}$$

$$= \frac{p(z_{t+1}|x_{t+1})p(x_{t+1}|x_t)}{p(z_{t+1}|z_{1:t})} p(x_{0:t}|z_{1:t}) \quad (11)$$

If we are interested in the iterative distribution the equation becomes:

$$p(x_{t+1}|z_{1:t+1}) = \frac{p(z_{t+1}|x_{t+1})p(x_{t+1}|z_{1:t})}{p(z_{t+1}|z_{1:t})} \quad (12)$$

$$= \frac{p(z_{t+1}|x_{t+1})p(x_{t+1}|x_t)}{p(z_{t+1}|z_{1:t})} p(x_{0:t}|z_{1:t}) \quad (13)$$

$$= \frac{p(z_{t+1}|x_{t+1}) \int p(x_{t+1}|x_t)p(x_t|z_{1:t})dx_t}{\int p(z_{t+1}|z_{1:t},x_{t+1})p(x_{t+1}|z_{1:t})dx_{t+1}} \quad (14)$$

$$= \eta\, p(z_{t+1}|x_{t+1}) \int p(x_{t+1}|x_t)p(x_t|z_{1:t})dx_t \quad (15)$$

Where η is a normalization factor to ensure that equation (15) represents a probability density function. Usually, the evaluation of this equation takes place in two steps. In the prediction phase, the x_{t+1} state is calculated starting from the x_t state, through the application of the transition model. Subsequently, the z_t observation is incorporated into the previously calculated probability density function, through the updating phase. The relationship between the present state and the prior state is depicted by the motion model. The location of the mobile robot in a global coordinate frame is defined as the state vector; $X_k = [x_k, y_k, \Theta_k]^T$, where, k indicates the sampling moment. The probability distribution is assumed to be Gaussian so that the state variable is fully described by a 3 * 3 covariance matrix P_k and the state expected value $\hat{x}_k$ (Mean estimated value). The control input u_k represents the motion command that moves the mobile robot from step k to step k + 1, since u_k can be expressed as; $u_k = [d_k, \Delta\Theta_k]^T$ denotes translation by distance d_k followed by a rotation angle $\Delta\Theta_k$. The state vector is calculated using the current state vector and the current control input by the state transition function f(.); $x_{k+1} = f(x_k, u_k, V_k)$. Since $V_k = [v_{1,k}, v_{2,k}]^T$ depicts process noise (also Gaussian with zero mean) $E(V_k) = [0,0]^T$, according to equations (1), (2), and (3) the state transition function is obtained.

$$f(x_k, u_k, V_k) =$$

$$\begin{pmatrix} x_k + (d_k + v_{1,k}).\cos(\Theta_k + \Delta\Theta_k + v_{2,k}) \\ y_k + (d_k + v_{1,k}).sin(\Theta_k + \Delta\Theta_k + v_{2,k}) \\ \Theta_k + \Delta\Theta_k + v_{2,k} \end{pmatrix} \quad (16)$$

Two independent error sources, translational and angular, were used to model the process noise covariance Q_k. The expression for Q_k is:

$$Q_k = \begin{pmatrix} \sigma_D^2 & 0 \\ 0 & \Delta\Theta_k^2 \sigma_{\Delta\Theta}^2 \end{pmatrix} \quad (17)$$

Where σ_D^2 and $\sigma_{\Delta\Theta}^2$ are the variances of translation d_k and rotation $\Delta\Theta_k$ respectively.

B. Measurement Model

Measurement function $h_i(X, p_i)$ determines the distance between the obstacles and robot's l RPLIDAR (Fig. 3).

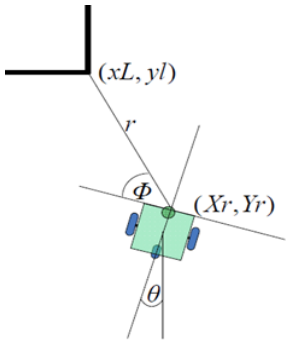


Fig. 3. Robot Pose from the Landmark.

In the world model, $p_i = (x_i, y_i)$ represents the point (occupied cell identified by the LIDAR). The LIDAR model utilizes distance readings that are linked to the causative obstacle.

$$z_{i,k} = h_i(X_k, p_i) + w_{i,k} \quad (18)$$

The measurement noise is represented by $w_{i,k}$. Because all distance measurements are utilized in parallel for the distance measurement value, the distance measurements $z_{i,k}$ are a single measurement vector. And r_k^i and $\emptyset_k^i$ components of the measurement form a diagonal matrix R_k.

C. EKF-based Pose Tracking

This paper describes EKF, as sensor fusion-based method for robot posture tracking. More a thorough discussion of the EKF localization method can be found in [22]. At time k the values of the control input vector u_{k-1} generated by wheel encoder are supplied to the equations that will be mentioned in this section, and the first-time update is performed to get the prediction estimate, and when fresh LIDAR measurements are received, these predictions are adjusted. The prior mean $\hat{X}_k^-$ is calculated by using the nonlinear Odometry function to propagate the predicted state.

$$\hat{X}_k^- = f(\hat{X}_{k-1}, u_{k-1}, E\{V_{k-1}\}) \quad (19)$$

The anticipated state covariance P_k^- is calculated by propagating the state covariance through a linearized system form.

$$P_k^- = \nabla f_X P_{k-1} \nabla f_X^T + \nabla f_u Q_k \nabla f_u^T \quad (20)$$

$\nabla f_X = \nabla f_X(\hat{X}_{k-1}, u_{k-1}, E\{V_{k-1}\})$ denotes the Jacobian of function f(.) with respect to the state X_k which can be computed as follows.

$$\nabla f_X = \begin{pmatrix} 1 & 0 & -(d_k + v_{1,k}) sin(\Theta_k + \Delta\Theta_k + v_{2,k}) \\ 0 & 1 & (d_k + v_{1,k}) \cos(\Theta_k + \Delta\Theta_k + v_{2,k}) \\ 0 & 0 & 1 \end{pmatrix} \quad (21)$$

$\nabla f_u = \nabla f_u(\hat{X}_{k-1}, u_{k-1}, E\{V_{k-1}\})$ is the Jacobian of $f(.)$ with respect to the control input u. It is important to note that when (9) and (10) are employed, the mean and covariance are only true to the first order of the associated Taylor expansion [23]. If no fresh LIDAR measurements are available at time k, or if all are discarded, there is no measurement update, and the estimated mean and covariance are assigned the anticipated values.

$$\begin{cases} \hat{x}_k = \hat{x}_k^- \\ P_k = P_k^- \end{cases} \quad (22)$$

Otherwise, a measurement update occurs, in which the initial forecasts of the approved LIDAR readings in $\hat{z}_{i,k}^-$ with the i-th component are as follows:

$$\hat{z}_{i,k}^- = h(\hat{X}_k^-, p_i) + E\{w_{i,k}\} \quad (23)$$

In time step k, the state estimate and covariance are calculated as follows:

$$\begin{cases} \hat{X}_k = \hat{X}_k^- + K_k(z_k - \hat{z}_k^-) \\ P_k = (I - K_k \nabla h_X) P_k^- \end{cases} \quad (24)$$

Since Z_k reflects actual LIDAR readings, where ∇h_X is the Jacobian matrix and we can obtain by calculating the derivative of measurement function with respect to state X_k:

$$\nabla h_X = \frac{\delta h_i(X_k, p_i)}{\delta X_k} = \begin{pmatrix} \frac{\delta r_k^i}{\delta \hat{X}_{k,x}} & \frac{\delta r_k^i}{\delta \hat{X}_{k,y}} & \frac{\delta r_k^i}{\delta \hat{X}_{k,\Theta}} \\ \frac{\delta \emptyset_k^i}{\delta \hat{X}_{k,x}} & \frac{\delta \emptyset_k^i}{\delta \hat{X}_{k,y}} & \frac{\delta \emptyset_k^i}{\delta \hat{X}_{k,\Theta}} \end{pmatrix} = \begin{pmatrix} -\frac{p_{i,x} - \hat{x}_{k,x}}{\sqrt{q}} & -\frac{p_{i,y} - \hat{x}_{k,y}}{\sqrt{q}} & 0 \\ \frac{p_{i,y} - \hat{x}_{k,y}}{q} & -\frac{p_{i,x} - \hat{x}_{k,x}}{q} & -1 \end{pmatrix} \quad (25)$$

Since:

$$q = (p_{i,x} - \hat{x}_{k,x})^2 + (p_{i,y} - \hat{x}_{k,y})^2 \quad (26)$$

And K_k is the Kalman filter which can be computed as follows:

$$K_k = P_k^- \nabla h_X^T (\nabla h_X P_k^- \nabla h_X^T + R_k)^{-1} \quad (27)$$

By implementation of previous equations, we can see the diverges in robot location in case of using Kalman filter (blue color) compared with sample Odometry (red color), as shown in Fig. 4.

After we have seen the diverges in robot pose using kalam filter compared with sample odmetry and deriving the necessary equations, We will go further, to see the effect of using kalman filter in reducing estimated error the following figure (Fig. 5) depicts the implementation of EKF on a differential drive mobile robot (turtelbot) using ROS.

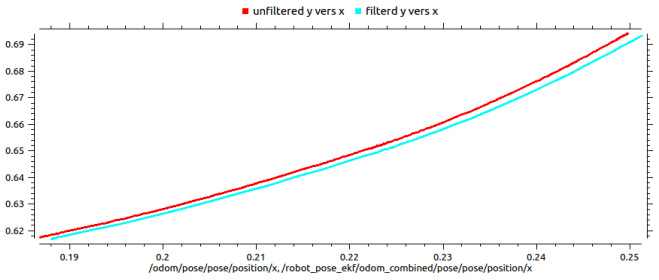


Fig. 4. Robot Ypos vers Xpos.

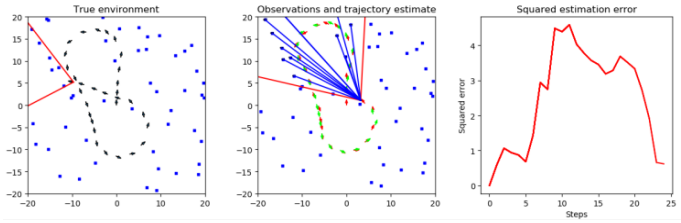
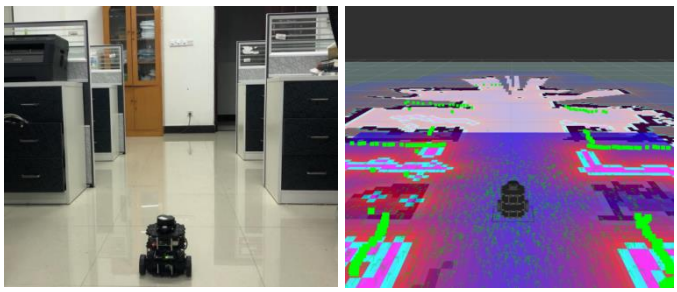


Fig. 5. True Robot Trajectory (Left), Estimated Trajectory (Middle) and Squared Positional Error (Right) using EKF Algorithm.

V. Discussion

In Fig. 5 estimates after the prediction step are shown as red arrows; estimates after the update step are shown as green arrows. Blue dots represent (static) landmarks and the red lines indicate the robot's field of view. Robot state covariance increases during the prediction step and decreases during the update step. The algorithm benefits more from the correction step when the motion covariance is large, as otherwise the prediction step already yields fairly precise estimates. A larger number of landmarks result in a more precise position estimate. Loop closing, e.g., observing landmarks that have already been observed at an earlier stage greatly decreases the covariance of both the robot state and landmark position. This effect is clearly visible in Fig. 5, where loop closure occurs twice: first around step 9 and then around step 20. Both times, the squared estimation error decreased. It's worth noting that the findings displayed above were obtained using ROS in simulated settings for obstacles, LIDAR measurements, linear and angular velocities; hence, with ROS, we can safely utilize the same software for actual robots, as shown in Fig. 6.



(a) (b)

Fig. 6. (a) Robot's Environment in Laboratory (b) Green Squared Dots Represent LIDAR Measurements Readings. Turtelbot Real EKF Implementation under ROS Robot Operating System.

VI. Conclusion

Extended Kalman filters are developed and tested as mobile robot posture monitoring methods. Odometry and sensor fusion LIDAR-based sensors are shown to significantly improve mobile robot localization, and the experimental results reveal substantial differences when utilizing sample Odometry and the EKF method. The algorithm is tested with different motion and measurement noise covariance, generated trajectories and a varying number of landmarks. This technique produces significant results in a limited environment and with a small number of Landmarks. However, it is clear that this approach is insufficient in the case of a large number of Landmarks, which would definitely result in an increase in the size of the state space.

Acknowledgment

This work was supported in part by the Guangxi Innovation Driven Development Science and Technology Project (grant # Guike AA18118002-3), in part by the Guangxi Key Lab of Manufacturing System and Advanced Manufacturing Technology (grant # 19-050-44-005Z), in part by Guangxi Science and Technology Base Talent Project (grant # Guike-AD19245141), and in part by Guangxi Natural Science Foundation Program (grant # 2020GXNSF BA297157).

References

[1] Borenstein, J. & Feng, L. (1996). Measurement and correction of systematic Odometry errors in mobile robots. *IEEE Transactions in Robotics and Automation*, 12(2), 869-880.

[2] C. Tarin Sauer, H. Brugger, E.P. Hofer & B. Tibken.(2002). Odometry error correction by sensor fusion for autonomous mobile robot navigation. *IEEE Instrumentation and Measurement Technology Conference*, Cat. No.01CH 37188. August 07, Budapest, Hungary.

[3] Digiampaolo E. & Martinelli F. (2014). Mobile robot localization using the phase of passive UHF RFID signals. *IEEE Transactions on Industrial Electronics*, 61(1), 365–376. doi:10.1109/TIE.2013.

[4] Leonardo Marín, Marina Vallés, Ángel Soriano, Ángel Valera & Pedro Albertos (2013). Multi Sensor Fusion Framework for Indoor-Outdoor Localization of Limited Resource Mobile Robots. *Sensor*, 13(10), 14133-14160. doi: 10.3390/s131014133.

[5] Iraj Hassanzadeh & Mehdi Abedinpour (2008). Design of Augmented Extended Kalman Filter for Differential Drive Mobile Robots. *Journal of Applied Sciences*, 8(16), 2901-2906.

[6] Jaeyong Park, Sukgyu Lee & Joohyun, Park (2009). Correction Robot pose for SLAM based on Extended Kalman Filter in a Rough Surface Environment. *International Journal of Advanced Robotic Systems*, 6(2), 67-72.

[7] Edouard Ivanjko, Mario Vasak, & Ivan Petrovic (2005). Kalman Filter Theory Based Mobile Robot Pose Tracking Using Occupancy Grid Maps. *International Conference on Control and Automation (ICCA2005)*, June 27-29, Budapest, Hungary.

[8] Yusuke Misono & Yoshitaka Goto (2007). Development of Laser Rangefinder-based SLAM Algorithm for Mobile Robot Navigation. *AICE Annual Conference*, September 17-20, Kagwa University, Japan.

[9] Angga Rusdinar, Jungmin Kim & Sungshin Kim (2010). Error Pose Correction of Mobile Robot for SLAM Problem using Laser Range Finder Based on Particle Filter. I*nternational Conference on Control, Automation and Systems*, Oct. 27-30, Gyeonggi-do, Korea.

[10] Mohammed Faisal & Mansour Alsulaiman (2016). Enhancement of mobile robot localization using extended Kalman filter. *Advances in Mechanical Engineering*, 8(11), 1–11. DOI: 10.1177/1687814016680142.

[11] Yi-Xiang Wang & Ching-Lung Chang (2020). ROS-base Multi-Sensor Fusion for Accuracy Positioning and SLAM System. *International Symposium on Community-centric Systems.* Tokyo, Japan DOI**:** 10.1109/CcS49175.2020.9231442.

[12] Zanwu Xia & Si Tang (2019). Robust self-localization system based on multi-sensor information fusion in city environments. *International Conference on Information Technology and Computer Application (ITCA)*. DOI: 10.1109/ITCA49981.2019.00011.

[13] Ivanjko, E., Petrovi, I. & Peric, N. (2003). An approach to Odometry calibration of differential drive mobile robot. *International conference of electrical drives and power electronics*. Sept. 24-26, pp. 519-523.

[14] Libal, U. & Płaskonka, J. (2014). Noise sensitivity of selected kinematic path following controllers for a unicycle. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 62(1), 3-13. *DOI: 10.2478/bpasts-2014-0001*.

[15] Kozłowski, K. & Michałek M. (2012). Trajectory tracking control with obstacle avoidance capability for unicycle-like mobile robot. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 60(3), 537-546. DOI: 10.2478/v10175-012-0066-x.

[16] Kozłowski, K. & Michałek M. (2019). Trajectory tracking and collision avoidance for the formation of two-wheeled mobile robots. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 20(19), 5915-924. DOI: 10.24425/bpas.2019.128652.

[17] Frank E. Schneider & Dennis Wildermuth (2004). Using Extended Kalman Fiter for Relative Localization in a Moving Robot Formation. *International workshop on robot Motion and Control*, June 17-20.

[18] NAGARAJAN M., BOSCO, J. & KANNAN R. (2017). Implementation of extended Kalman filter-based simultaneous localization and mapping: a point feature approach, *Sādhanā*, 42(9), 1495–1504. *doi: 10.1007/s12046-017-0692-y*.

[19] Xifeng Li & Yongle Xie (2013). State Estimation Based on Generalized Gaussian Distributions, *Metrology and Measurement Systems*, 1, 65-76. *DOI: 10.2478/mms-2013-0006*.

[20] Jonas Skeivalas, Eimuntas Paršeliūnas, Raimundas Putrimas & Dominykas Šlikas (2020). An influence of the covariance between single orbit parameters on the accuracy of observations of the pseudo ranges and phase differences. *Metrology and Measurement Systems*, 27(1), 131-140. *DOI: 10.24425/mms.2020.131721*.

[21] Jeremy Kolansky & Corina Sandu (2013). Real-time parameter estimation study for inertia properties of ground vehicles. *Archive of Mechanical Engineering*, 60, 17-21. DOI: 10.2478/meceng-2013-0001

[22] Heonmoo Kim & Yosoon Choi (2020). Comparison of Three Location Estimation Methods of an Autonomous Driving Robot for Underground Mines. *Applied Sciences*, 10, 4831. doi:10.3390/s131014133.

[23] Housein, A. & Gao Xingyu (2021). Simultaneous Localization and Mapping using differential drive mobile robot under ROS, *Intelligent Manufacturing and Automation Technology (MEMAT)*, January 15-17, Guilin, China. doi:10.1088/1742 6596/1820/1/012015.

The Adoption of Digital Games Among Older Adults

Nurul Farinah Mohsin[1], Suriati Khartini Jali[2], Sylvester Arnab[3], Mohamad Imran Bandan[4], Minhua Ma[5]
Faculty of Computer Science and Information, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia[1, 2, 4]
Disruptive Media Learning, Coventry University, United Kingdom[3]
Falmouth University, Cornwall, United Kingdom[5]

***Abstract*—The revolution of technology brings many benefits towards diverse population. Digital game is one of the digital technologies that has potential to facilitate older adults' daily routine. However, some of them faces challenges to adopt the usage of digital games in their daily lives, one of which is that most commercial games are not suitable for older people. This paper discusses the investigation into the challenges associated with the older adults' adoption of digital games, their interaction, and experiences with digital games and specifically explores the andragogical perspectives, and game design attributes. A set of questionnaires consisted of open-ended and close-ended questions were distributed, targeting the older adults across Malaysia, using online and non-probability sampling technique. 81 respondents were recruited, and 56 respondents (n=56) were eligible in this study. Four participants were recruited for informal interview session. The analysis of the results indicates that the older adults' perception of digital games and game design aspects are the major factors influencing their digital game adoption. Game designs are important to attract many older adults to experience and interact with digital games.**

***Keywords*—Digital games; Malaysia; older adults; technology**

I. Introduction

The number of older adults in Malaysia has increased from 3.4 million in 2019 to 3.5 million in 2020, and by 2030, more than 15% of the Malaysian population will be the older population [1]. [2] stated that in Malaysia older adults aged 60 and over are categorised based on the United Nations (UN) age capping. Two main factors that contribute to the increasing number of the older population in Malaysia are lower birth rates and the declining Total Fertility Rate (TFR) [3]. Older adults tend to experience negative ageing effects such as declining in cognitive abilities and physical abilities. The advancement of digital technology such as digital games can help to facilitate the negative ageing. Digital game technology has benefits beyond the purpose of enjoyment, where it has increasingly been applied as a tool for psychological, cognitive, and neuropsychological rehabilitation for older adults [4]. Despite the benefits of digital game technology for the older population, certain older adults may confront difficulties and challenges when it comes to experiencing and participating in it.

The main problem is the game design is not suitable for older adults and does not consider their incapacity. Most of the commercialized digital games are design and develop for the general type of game and aiming for the younger user in mind.

Games specifically designed for older adults are not commercially available yet [5]. Some of the exergames can be used as a tool for exercising however, older adults who experience a decline in physical abilities might not be able to obtain a complete gaming experience due to limited physical movements. Some digital games were designed with a complex interface. Older adults who experience cognitive issues might have difficulties interacting with these games.

This paper discusses the challenges faces by the older adults aged 55 to 75 within the Malaysian context, their gaming experiences and interactions with digital games, and the associated influencing factors associated with the andragogical perspectives, and game design attributes. There are two research questions aimed to be addressed in this study:

Research question 1: What are the challenges related to older adults' interaction with digital game technology that needs to be considered?

Research question 2: How can their gaming experiences inform design considerations?

II. Literature Review

Various studies have highlighted challenges associated with older adults' experience of digital games. One of the critical factors that often influence their engagement with digital technology is the psychological factor. For instance, the fear of their inability to use technologies correctly often affects their confidence and level of readiness and acceptance [6]. Blažič and Blažič [7] suggested the root of the problem is the gap in knowledge on how to use digital devices effectively. Other factors include the natural ageing characteristics such as the decline in cognitive and physical abilities, which was not considered during the design of technologies such as games.

Game designs are classified as mechanics, dynamics, and aesthetics. Game mechanics and dynamics design refer to the interaction aspect and how the game operates. Aesthetics, on the other hand, covers the interface design in the game, the look and feel as perceived by the player. Garcia et al. [8] revealed in their study that participants have difficulties remembering certain features in the game and are confused with the game interface that presents too many options. It is essential for digital games designed for older players to have a user-friendly interface, on easy-to-use platform, and are simple to learn [9]. The number of researchers that design and develop games for older adults is still relatively limited. As stipulated by [10], most games do not consider older adults' needs and interests. There is also a lack of research that correlates the attributes of game technology with andragogical considerations and challenges faced by the target group.

To overcome the challenges, it is crucial to embed andragogical perspectives while designing and developing the

digital game for older adults to help them throughout the learning process. Andragogy is the art and science helping adult learners. There are six fundamental principles in andragogy [11], (1) learner's need to know, (2) self-concept of the learner, (3) prior experience of the learner, (4) readiness to learn, (5) orientation of learning and (6) motivation to learn. According to [12], the process of learning something new for older adults would be more accessible when they know the benefits and relevance of the technology in their daily life. They have conducted a study with older adults aged 55 and above in Coventry, United Kingdom. Based on their findings, the participants learned new knowledge when they engaged with the digital games, they thought were beneficial. In this study, the adoption of digital game among older adults in Malaysia were considered to construct a framework for designing and developing game for older adults using andragogical perspectives.

The purchasing power of the elderly consumers is growing and is expected to rise rapidly in the next decade due to the ageing population globally. Leisure and entertainment, such as digital game, form a considerable share of an average senior's budget. The findings of this study will inform and help game designers developing products for this fast-growing market.

III. Materials and Methods

This section discusses the materials and methodology used to conduct this study. The recruitment process of participants for this study will be discussed in Section A, instruments used will be explained in Section B and data collection and analysis procedures will be described in Section C. Fig. 1 depicts the flow of this study.

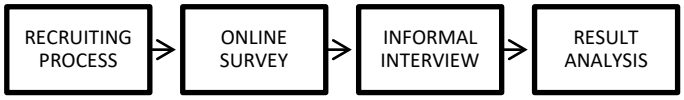


Fig. 1. Study Flow.

A. Participants

This study targeted Malaysians between 55 and 75 years old who resided in Malaysia during the study period, regardless of their experience in digital games. The recruitment of participants used convenience and snowball sampling techniques. An online survey link was distributed and disseminated using social media platforms (e.g., Whatsapp, Facebook, Telegram, Instagram, Twitter, Microsoft Outlook). All eligible participants provided informed consent online before completing the survey administered via Google Form.

B. Instruments

The study implemented a set of questionnaires that consisted of open- and close-ended questions categorised under six key sections: (1) participant background, (2) technology usage, (3) game experience, (4) game usability, (5) challenges, and game recommendation, and (6) opinion and suggestion. The questionnaire was disseminated to the target participants via an online survey (Google Form). All participants were required to complete Sections 1 and 2. Subsequently, participants who indicated experience in digital games were given access to Section 3-5 in the questionnaire.

Section 1 Participant background was used to gather the demographic background of the respondents. Section 2 contains questions on technology use, which were used to ascertain participants' computer abilities and familiarity with digital games. Those who have experience with digital games then answered Sections 3 and 4, which are used to assess the game experience and the usability of the game played by the respondents. The game experience was measured by using Game Experience Questionnaire (GEQ) [13], and the game usability was measured by using the System Usability Scale (SUS) [14]. The GEQ was adopted from [15].

Section 5 asked the participants to discuss their challenges and recommendations on game design based on their prior gaming experience. Finally, in Section 6, participants could give their opinions on the usage of the digital game among older adults in Malaysia.

C. Data Collection and Analysis Procedure

All data collected via Google Form was stored in Microsoft Excel. The responses from the participants were analysed quantitatively. The Statistical Package for the Social Sciences (SPSS) software (version 22; SPSS Inc., Chicago, IL, USA) was used to assist the researchers in analysing the data. Two types of statistical analysis were used in this study. Descriptive analysis was used to analyse the participants' background, and Chi-Square was used to test the association between participants' computer skills and their experience playing digital games. Responses from the open-ended questionnaire were discussed qualitatively. Informal interview sessions with four participants were conducted. The sessions were recorded using audio recorder and data were analysed.

IV. Results

In this section, results obtained from the questionnaire were discussed. Section A addresses the participants' backgrounds. Section B discusses the participants' technology use. Section C explains the participants' computer abilities and experience with digital games. Finally, section 4.4 highlights the obstacles that may affect older people's digital game use.

In this study, 81 responses were recorded, however, due to the targeted age group of 55 to 75 years old for this study, we excluded 25 responses as they are not in the age range. Based on the data obtained, 41 of the participants were between the age of 55-60 years old, followed by an age range between 61-65 years old (n=9), 66-70 years old (n=5), and 71-75 years old (n=1). 33 of the participants were female and 23 were males. Based on the participants' marital status, 45 of them were married, 9 were widowed, 1 was single, and 1 was divorced. In terms of employment status, 20 of the participants were full-time workers, 20 retired, 13 were unemployed, and 3 were in part-time employment. All respondents were recruited from Malaysia, where 24 of them are from Sabah, followed by Sarawak (n=14), Wilayah Persekutuan (n=4), Selangor (n=4), Pahang (n=3), Melaka (n=2), Perak (n=1), Johor (n=1), and Perlis (n=1).

Informal interview session was involved with four participants from different districts and state in Malaysia. 2 female participants are from Kota Kinabalu Sabah, while remaining 2 male participants are from Labuan and Perak respectively.

A. Technologies usage among Participants

This section of the questionnaire required respondents to give information on their technology usage. Participants were required to state their computer skills, type of devices owned, and their experience using digital games. Based on the result obtained, 4 participants rated their computer skills as 'expert', 17 rated 'competent', 24 rated 'novice', and 11 of them rated themselves as none. In response to the question: "What kind of devices you own", most of the participants owned at least one type of digital device. 41 of the participants owned a smartphone.

The focus of this study is to investigate their game experience and game consideration. Based on the responses obtained, only 10 of the respondents had experience in using digital games. Table I presents the game name and game type played by the respondents. Most participants who indicated 'No' for their experience in using digital game stated that playing digital games is a waste of time. They mentioned how they are not interested in playing and thought that playing digital games is irrelevant for older populations.

B. Relationship between Computer Skills, Employment Status, and Experience with Digital Games

Table II shows the crosstabulation between participants' computer skills, employment status, and their experience in playing digital games. Based on the result obtained, those with experience playing digital games were novice and competent participants from various employment statuses. The skills and employment status were not influenced the participants' experience in playing digital games.

TABLE I. GAME NAME AND GAME TYPE PLAYED BY THE RESPONDENTS

Game Name	Game Type	Frequency (n)	Percentage (%)
Candy crush	Puzzle	7	70
Word Search	Puzzle	3	30
Bubble	Puzzle	2	20
Toybear	Puzzle	1	10
Bubblesoap	Action	1	10
Deer Hunter	Simulation	1	10
Galaxy Force	Third Person Shooter Game	1	10
Player Unknown Battle Ground (PUBG)	First Person Shooter Game	1	10
Mobile Legend (ML)	Adventure	1	10
Farm Ville	Simulation	1	10
Angry Bird	Puzzle	1	10
Mario	Adventure	1	10
Word Puzzle	Puzzle	1	10

TABLE II. CROSS TABULATED RESULTS OF SKILLS, EMPLOYMENTS STATUS, AND RESPONDENTS' EXPERIENCE PLAYING WITH DIGITAL GAMES

Skills			Experience Digital Game		Total
			Yes	*No*	
Expert	Employment Status	Full time	0	2	2
		Retired	0	2	2
	Total		0	4	4
Competent	Employment Status	Full time	2	7	9
		Part time	0	2	2
		Retired	2	4	6
	Total		4	13	17
Novice	Employment Status	Full time	2	6	8
		Part time	0	1	1
		Retired	1	8	9
		Unemployed	3	3	6
	Total		6	18	24
None	Employment Status	Full time	0	1	1
		Retired	0	3	3
		Unemployed	0	7	7
	Total		0	11	11

A Chi-Square test was conducted to test the association between participants' computer skills and their experience playing the digital game. Table III shows the results of the Chi-Square test between computer skills and participants' experience playing digital games. Results revealed that there is no association between computer skills and experience playing digital games, $\chi 2 = 4.469a$, $p = .215$. Thus, there is no association found between computer skills and gameplay experience among the participants. Based on the findings, the level of participants' computer literacy is not a major factor that affects their engagement with digital games as those who considered themselves as competent or novice were more engaged with digital games than those who considered themselves as an expert.

TABLE III. CHI-SQUARE TESTS

	Value	df	p-value
Pearson Chi-Square	4.469a	3	.215
Likelihood Ratio	7.010	3	.072
Linear-by-Linear Association	.372	1	.542
N of Valid Cases	56		
a. 5 cells (62.5%) have expected count less than 5. The minimum expected count is .71.			

C. Challenges that may Impact Older Adults' used of Digital Games

In this section, responses from participants who indicated have experience with digital games are discussed based on the questions in Sections 3 to 6 in the questionnaire. There were two research questions related to these sections.

Research Question 1: What are the challenges related to the older adults' interaction with digital game technology that needs to be considered?

Based on our findings, the game design played by the respondents was one of the challenges for them to interact with digital game technology. Older adults might be affected by negative ageing impacts, for example, the decline in physical and cognitive abilities. Participant 18, said, "tulisan kecil" (small wordings) is one of the challenges for her to adapt to digital game technology. Apart from that, the screen size of devices used during the gameplay session was small and it affected their gameplay session. According to the principles of andragogy, older adults must have self-concept and experience with the platform used to play the digital game so that they can fully control the gaming experience.

Besides game design, the perception of older adults is one of the challenges to attract them to play the digital game. In Section 6 of the questionnaire, participants can give their opinion regarding the usage of digital games among older adults. Based on the findings, most of the participants find digital games irrelevant for older populations and perceive playing games as a waste of time. Also, some of them are not interested in playing digital games and feel that digital game interaction is complicated and challenging. One of the respondents commented, "kurang memberi faedah" (give fewer benefits) on question asking opinions regarding the usage of digital games among older adults. Despite the generally negative views on games, one of the respondents stated that digital games can be used as a tool to relieve stress, but older adults need to get proper guidance on how to use them as some older adults are not familiar with digital games.

Research Question 2: How can their game experiences inform design consideration?

A reliability test was conducted to identify the internal consistency and specified measurement's usability. Cronbach's Alpha defines the internal consistency or average correlation of items in the GEQ. The Cronbach's Alpha value is 0.820 where it is acceptable and reliable [16]. GEQ is a well-established questionnaire used in many previous studies [17, 18, 19, 20]. Santos et al. [17] used GEQ to report their participants' experience during game play and their sense of social presence.

As mentioned earlier, only 10 out of 56 participants had experience in engaging with digital games. Game Experience Questionnaire (GEQ) was used to measure their level of experience during gameplay sessions [15]. We used the 5-Likert scale (strongly disagree, disagree, neutral, agree, strongly agree). GEQ consists of 33 items, and they measure 7 players' experiences include immersion, flow, competence, tension, challenge, positive affect, and negative affect during gameplay sessions.

Table IV presents the results for positive affect items in the GEQ. Positive items measure players' emotional experience (e.g., enjoyment). The results show that most of the participants agreed that the gaming experience brings a positive effect on them. Cronbach's Alpha value for positive items is 0.957. The questions "I thought it was fun" and "I felt happy" have the highest mean values (M=3.90). This indicates that the respondents' gaming experience was fun, and they felt happy during gameplay. The mean for the question "I felt content" has the lowest mean value (M=3.20) and it might be influenced by the theme of the games played by the participants. This question linked to the andragogy principle, where the participants were not content during the gameplay session. Based on our findings, we found that the participants who respond neutrally in this question played puzzle game type. Each participant played a different type of digital game, where their emotional feeling towards the gaming experience might be different from others.

The second item measured in GEQ is competence and the Cronbach's Alpha value is 0.833. Based on the findings, most of the participants' scores were neutral except for the question "I felt successful" where they agreed based on their gaming experience. However, one of the participants strongly disagreed with the question "I felt successful". This result might be influenced by the game type played by participant which was Word Puzzle. The game requires greater effort to solve a puzzle than other puzzle games and if the player could not solve it, they might feel unsuccessful and frustrated. Table V illustrates the results for competence items. One of the participants during the interview session suggested to include the scoreboard elements in the game. He said the element can gives satisfaction during gameplay session as he can see the progress score when he plays the game.

TABLE IV. SUMMARY OF POSITIVE ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
I felt content	10	2	4	3	32	3.20	.632
I thought it was fun	10	3	4	4	39	3.90	.316
I felt happy	10	2	5	4	39	3.90	.876
I felt good	10	2	4	4	37	3.70	.675
I enjoyed it	10	2	5	4	37	3.70	.949
Valid N (listwise)	10						

TABLE V. SUMMARY OF COMPETENCE ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
I felt skillful	10	2	4	3	31	3.10	.568
I felt competent	10	3	4	3	34	3.40	.516
I was good at it	10	3	5	3	35	3.50	.707
I felt successful	10	1	5	4	36	3.60	1.075
I was fast at reaching the game's targets	10	z	4	3	31	3.10	.738
Valid N (listwise)	10						

TABLE VI. SUMMARY OF IMMERSION ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
I was interested in the game's storyline	10	2	5	4	34	3.40	1.075
It was aesthetically pleasing	10	1	5	3	32	3.20	1.033
I felt imaginative	10	2	4	3	29	2.90	.738
I felt that I could explore things	10	1	5	4	33	3.30	1.160
I found it impressive	10	1	5	4	34	3.40	1.174
It felt like a rich experience	10	2	4	3	32	3.20	.632
Valid N (listwise)	10						

Table VI shows the results obtained for the immersion aspect. There were six questions in the immersion item and the Cronbach's Alpha value is 0.904. Based on the findings, most of the participants were either neutral or agreeable on the immersion questions on their previous gaming experience.

There are five questions related to flow items. Table VII depicts the results from preliminary analysis for flow items. The Cronbach's Alpha value for flow items is 0.823. A majority of the participants agreed on the "I forgot everything around me" and "I was deeply concentrated in the game" items. This indicates the participants were focused during the gameplay session.

The results obtained from the preliminary analysis of negative affect during the gameplay experience are shown in Table VIII Cronbach's Alpha value for negative items is 0.821. Most of the participants disagreed with the negative effect during gameplay, such as giving them a bad mood or feeling bored. For the aspects related to "I thought about other things" and "I found it tiresome", most of the participants responded neutrally; perhaps influenced by the type of game that they played.

One of the participants in the interview session said he feels frustrated during the gameplay session when he cannot reach or complete the task in game after few times attempt. Furthermore, he said this kind of frustration can lead to decrease the level of desire to proceed with gameplay session.

There are five items measured in challenges as listed in Table IX. The Cronbach's Alpha value is 0.844. Based on the results obtained, most of the respondents disagreed with the question "I felt pressured" and "I felt time-pressured". These questions reflected the games played by the respondents, where most of the games played by them do not have a time limit to complete the task.

During the interview session, participants were asked regarding their opinion on digital games and the features they want in the game. Some of the like challenges, elements in the game, there it gives satisfaction to them during gameplay session. However, one of the participants said she preferred an easy game where it does not require much effort to play and control the game and with no complicated instruction to learn and understand.

TABLE VII. SUMMARY OF FLOW ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
I was fully occupied with the game	10	2	5	3	32	3.20	.919
I forgot everything around me	10	2	4	4	32	3.20	.919
I lost track of time	10	2	5	2	31	3.10	1.101
I was deeply concentrated in the game	10	1	5	4	37	3.70	1.160
I lost connection with the outside world	10	1	5	2	26	2.60	1.174
Valid N (listwise)	10						

TABLE VIII. SUMMARY OF NEGATIVE ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
It gave me a bad mood	10	1	5	2	23	2.30	1.160
I thought about other things	10	1	4	3	30	3.00	.816
I found it tiresome	10	1	4	3	29	2.90	.876
I felt bored	10	2	3	2	24	2.40	.516
Valid N (listwise)	10						

TABLE IX. SUMMARY OF CHALLENGE ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
I thought it was hard	10	1	5	3	30	3.00	1.155
I felt pressured	10	2	5	2	29	2.90	1.287
I felt challenged	10	2	5	3	34	3.40	1.075
I felt time pressure	10	2	4	2	27	2.70	.823
I had to put a lot of effort into it	10	2	5	3	32	3.20	.919
Valid N (listwise)	10						

TABLE X. SUMMARY OF TENSION ITEMS

	N	Min	Max	Mode	Sum	Mean	Std. Deviation
I felt annoyed	10	1	4	2	24	2.40	.843
I felt irritable	10	1	4	2	25	2.50	.972
I felt frustrated	10	1	4	2	26	2.60	1.075
Valid N (listwise)	10						

It can be seen from findings shown in Table X that most of the participants disagreed on the tension items. Digital games are often known as a tool that can bring enjoyment to the players. One of the participants in this study agreed that the gaming experience made him annoyed, irritable, and frustrated. Based on the game name indicated by the participant, we found that he played Deer Hunter and Galaxy Force game. Both games played by the participant required him to achieve the target in the game. Thus, it might influence his feeling during the gameplay session. The Cronbach's Alpha value is 0.957.

The game evaluation used in this study is SUS, and the Cronbach's Alpha value is 0.539. The value is unacceptable as it is below 0.60. The reliability of SUS in this study is inconsistent as every user experienced a different type of user interface based on their previous gameplay session. The overall score for the SUS value is 56.25. This value is acceptable and falls in the marginal range for the SUS score [21] F-+ig. 2.

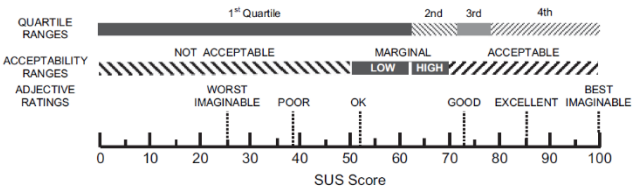


Fig. 2. SUS Adjective Rating [21].

V. DISCUSSION

From the survey results, only 18% of the total participants have experience in playing digital games. We found that the main factor that influences older adults to interact and experience digital games is their perception of digital games. It is important to ensure older adults acknowledge the benefits they gained from playing digital games. Apart from using it as an entertainment tool, digital games can be used as an innovative healthcare platform.

Computer skills do not affect the participants' experience with the digital game. Older adults who indicated their skills as competent and novice have had experience with digital games compared to those who indicated their skills as an expert. The game design is the main factor that influences the game experiences for older adults.

Based on the list of games played by the respondents, most of them were designed with no time pressure to help older adults focus on gameplay and distract them from feeling time pressure. The most popular game genre played by the participants in this study is puzzle, where it was designed with simple operation and highly achievable goals. Owing to older adults' decline in cognitive capacity, simple games may appeal to older adults. Apart from that, playing puzzle games does not require complex computing skills, making it more straightforward for older adults to play. Furthermore, puzzle game is a type of game that is familiar to this population, and it does not require them to learn new things, and the in-game instruction and rules are not complicated.

In our findings, most of the games played were puzzle games with a limited storyline for the player to immerse in the gaming experience and imagination may be limited during gameplay sessions. The participants played the CandyCrush game, where it is a part of a puzzle game. The game's mechanics and dynamics are repetitive, where it does not change and give the element of surprise to the players. As a result, they might feel tired and not enjoying the gameplay. It is important to design game with surprise elements to ensure the player enjoy throughout the gaming session.

The results for immersion category were influenced by the game's storyline, interface, and the mechanics of the game played by the respondents. It is important for the aesthetics, dynamics, and mechanics in the game should be designed to cope with older adults' cognitive and physical abilities. As indicated in the principle of andragogy, older adults will feel motivated to learn something if it can attract their attention and bring benefits to them.

Older adults will give their attention to the gameplay session if they understand the game flow, aligning with the key principle of andragogy, which are the learners' self-concept and their readiness to learn new things. The results revealed that the game played by the participants can attract their attention and make them concentrate during gameplay sessions.

VI. CONCLUSION

The findings in this paper are subject to at least three limitations. First, the number of participants eligible included in our study is relatively small. There are 81 responses recorded; however, only 56 responses are eligible to include in this study due to the age limitation. Other than the number of respondents, only ten participants have experience playing the digital games and the games they played were different. Thus, the gaming experience of the respondents might differ from other respondents. The verbal and non-verbal behavior of the participants during the gameplay session is not observed and the interview session with target people is not conducted due to the limitation to meet the respondents face-to-face during Movement Control Order (MCO) imposed in Malaysia. To overcome the limitations, it is recommended to conduct a face-to-face session with older adults to observe their verbal and non-verbal behaviour during gameplay sessions and use the same game type to measure their gaming experience. This paper has shown the data on the adoption of older adults to interact with digital technology and their opinions on the relevance of digital games towards the older population in Malaysia.

ACKNOWLEDGMENT

This research is fully supported by the Kementerian Pengajian Tinggi Malaysia, Fundamental Research Grant Scheme, FRGS/1/2019/ICT01/UNIMAS/03/1. The authors

fully acknowledged the Ministry of Higher Education (MOHE) and Universiti Malaysia Sarawak for the approved fund, which makes this important research viable and effective.

REFERENCES

[1] Department of Statistics, Current Population Estimates, no. December. Putrajaya, Malaysia: Department of Statistics, Malaysia, 2020.

[2] N. M. Yunus, N. H. Abd Manaf, A. Omar, N. Juhdi, M. A. Omar, and M. Salleh, "Determinants of healthcare utilisation among the elderly in Malaysia," Institutions Econ., vol. 9, no. 3, pp. 117–142, 2017.

[3] W. M. S. W. Ibrahim et al., "Population and Demographics: Ageing," vol. 1, pp. 1–2, 2017.

[4] F. Zhang and D. Kaufman, "Physical and Cognitive Impacts of Digital Games on Older Adults: A Meta-Analytic Review," J. Appl. Gerontol., vol. 35, no. 11, pp. 1189–1210, 2016.

[5] N. Khalili-Mahani et al., "For Whom the Games Toll: A Qualitative and Intergenerational Evaluation of What is Serious in Games for Older Adults," Comput. Games J., vol. 9, no. 2, pp. 221–244, Jun. 2020.

[6] T. Yeh, F. Pai, and M. Jeng, "The Factors Affecting Older Adults ' Intention toward Ongoing Participation in Virtual Reality Leisure Activities," Int. J. Environ. Res. Public Heal. Artic., vol. 16, no. 3, 2019.

[7] B. J. Blažič and A. J. Blažič, "Overcoming the digital divide with a modern approach to learning digital skills for the elderly adults," Educ. Inf. Technol., vol. 25, no. 1, pp. 259–279, 2019.

[8] J. A. Garcia, W. L. Raffe, and K. F. Navarro, "Assessing user engagement with a fall prevention game as an unsupervised exercise program for older people," ACM Int. Conf. Proceeding Ser., 2018.

[9] C. T. Lin and S. S. Chuang, "A Study of Digital Learning for Older Adults," J. Adult Dev., vol. 26, no. 2, pp. 149–160, 2019.

[10] X. Wang, X. Yao, and J. Gu, "Attraction and Addiction Factors of Online Games on Older Adults: A Qualitative Study," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11593 LNCS, pp. 256–266, 2019.

[11] C. Malliarakis, F. Tomos, O. Shabalina, and P. Mozelius, "Andragogy and e.M.o.t.i.o.n.: 7 key factors of successful serious games," in Proceedings of the 12th European Conference on Games Based Learning (ECGBL 2018) :, 2018, vol. 12, pp. 371–378.

[12] S. K. Jali and S. Arnab, "The Perspectives of Older People on Digital Gaming : Interactions with Console The Perspectives of Older People on Digital Gaming : Interactions with Console and Tablet-based Games," Vaz Carvalho C., Escudeiro P., Coelho A. Serious Games, Interact. Simulation. SGAMES 2016. Lect. Notes Inst. Comput. Sci. Soc. Informatics Telecommun. Eng., vol. 176, no. December, 2017.

[13] W. Ijsselsteijn et al., "Measuring the Experience of Digital Game Enjoyment," J. Pers., vol. 2008, no. June 2014, pp. 7–8, 2008.

[14] R. A. Grier, A. Bangor, P. Kortum, and S. C. Peres, "The system usability scale: Beyond standard usability testing," Proc. Hum. Factors Ergon. Soc., pp. 187–191, 2013.

[15] W. A. IJsselsteijn, Y. A. W. de Kort, and K. Poels, "The Game Experience Questionnaire," Eindhoven: Technische Universiteit Eindhoven, 2013.

[16] N. Niksadat, S. Rakhshanderou, R. Negarandeh, A. Ramezankhani, A. Vasheghani Farahani, and M. Ghaffari, "Development and Psychometric Evaluation of Andragogy-based Patient Education Questionnaire (APEQ)," Am. J. Heal. Educ., vol. 50, no. 6, pp. 390–397, 2019.

[17] L. H. Santos et al., "Pervasive game design to evaluate social interaction effects on levels of physical activity among older adults," J. Rehabil. Assist. Technol. Eng., vol. 6, no. June, p. 205566831984444, 2019.

[18] A. Barenbrock, M. Herrlich, and R. Malaka, "Design lessons from mainstream motion-based games for exergames for older adults," Conf. Proc. - 2014 IEEE Games, Media, Entertain. Conf. IEEE GEM 2014, app. 1–8, 2015.

[19] N. A. Merriman, E. Roudaia, M. Romagnoli, I. Orvieto, and F. N. Newell, "Acceptability of a custom-designed game, CityQuest, aimed at improving balance confidence and spatial cognition in fall-prone and healthy older adults," Behav. Inf. Technol., vol. 37, no. 6, pp. 538–557, 2018.

[20] J. Barbara, "Measuring user experience in board games," Int. J. Gaming Comput. Simulations, vol. 6, no. 1, pp. 64–79, 2014.

[21] A. Bangor et al., "An Empirical Evaluation of the System Usability Scale Usability Scale," vol. 7318, 2008.

Evaluation of Consumer Network Structure for Cosmetic Brands on Twitter

Yuzuki Kitajima[1]
Graduate School of Science and Engineering, Chuo University
Bunkyo-ku, Tokyo, Japan

Kohei Otake[2]
School of Information and Telecommunication Engineering
Tokai University, Minato-ku
Tokyo, Japan

Takashi Namatame[3]
Faculty of Science and Engineering
Chuo University, Bunkyo-ku
Tokyo, Japan

***Abstract*—Since the early 2000s, the Internet has become increasingly popular for the development of information dissemination technology and as a platform for interaction. Therefore, the penetration rate of Social Networking Services (SNSs) is also increasing. Using the accounts created on SNSs, companies can disseminate information and communicate with users on SNSs for marketing purposes. Moreover, there are several influencer marketing activities that use influencers who are highly influential in their surroundings as marketing using SNSs. In this study, we aim to identify influencers on Twitter and consumer network structures for six cosmetic brands. Specifically, create a consumer network for each of the six cosmetic brands using follower data obtained from Twitter is created to identify the network structure. Furthermore, brand influencers were also identified. The consumer network of all six cosmetic brands was created to identify the influencers in the cosmetics industry. We compared the influencers of the brands with the influencers of the entire industry to examine any differences.**

***Keywords*—Social networking services; community structure; network analysis; consumer network; influencer marketing**

I. Introduction

With the development of information technology and the widespread use of Internet devices such as PCs and smartphones, the number of users on social networking services (SNSs) has been increasing yearly. SNSs are a form of media that allow users to easily share and disseminate information in real-time. They also allow two-way communication between the company and the consumer. In particular, the number of SNSs users in Japan is diverse in age groups and is expected to grow to approximately 80% of the population by 2020 [1]. According to a Japanese market research firm, LINE, Twitter, and Instagram are the most popular SNSs tools in Japan. LINE is a messaging tool that serves as a short message service (SMS). In contrast, Twitter and Instagram are SNSs tools that allow people with similar interests to communicate with each other anonymously and closely, and they are popular worldwide. In particular, Twitter ranks 15th in the global survey, while it ranks 2nd in Japan, which indicates a high penetration rate in Japan [2].

The spread of SNSs is not only limited to consumers, but also affects various companies. As shown in Fig. 1, number of corporate SNSs accounts has increased by approximately 10% from approximately 30% in a year from 2017 to 2018 [3]. It is thought that the main reason for the increase in the number of corporate accounts on SNSs is that the diffusion of information by general users is expected to improve the product and brand recognition of companies. Considering the affinity between SNSs and marketing, we consider that marketing measures using SNSs are effective in industries with a large number of brands, such as the wholesale and retail industries. In addition, there are fashion and cosmetics industries where competition among brands is fierce, and it is assumed that these industries are also effective in improving brand recognition. Therefore, marketing using SNSs has become mainstream in recent years.

For example, Ferreira et al. [4] investigated the emergence of communities of co-commenters, that is, groups of users who often interact by commenting on the same posts and may be driving the ongoing online discussions. Their research used Instagram. The reason for this is that in recent years, social networking sites have been used as a source of information among young people, and politicians have also started to use this platform to spread information about political issues. In addition, Wang et al. [5] identified and tested the main factors related to SNS brand communities that can predict purchase intention. Their study suggested that companies should strategically manage consumers' SNS brand community experiences and commitment. Other theoretical implications and managerial implications were also discussed. This kind of research on customer behavior using social media is also conducted in the airline and hotel industries [6,7]. It is considered that the identification of influencers and consumer communities is an effective way to utilize SNSs in marketing. Influencers are people who have a large number of fans, especially SNSs, and have a large impact on the public. The identification of consumer communities is effective in identifying real customers of a brand.

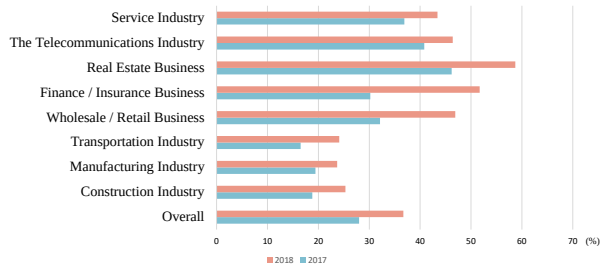


Fig. 1. Penetration of SNSs for each Industry.

In previous research on network communities, Girvan et al. [8] focused on properties found in many networks, such as small-world and power-degree distributions. Specifically, they focused on detecting community structures where nodes are tightly connected and there are loose connections between them, which is the case in many networks. As a result, they proposed a method for finding community boundaries using a centrality index and detected useful community partitions in the validation data. Chunaev [9] conducted a survey of social network analysis. Specifically, they aimed to investigate and clarify the situation of community detection in social networks with node attributes. They found that community detection methods that deal with both network structure and attributes are effective for social network analysis. Mizuno et al. [10] classified CtoC (customer to customer) interactions by digitalization into four types and reviewed the research trends of each type. In their review, they suggested that the closer the consumers were to each other, the easier it was for information to propagate on SNSs. They also suggested that it is important to know which consumers or targets to reach out to first in the social graph. Himelboim et al. [11] created a network of users based on the conversations on Twitter and discussed the network structure by information flow. Through their research, they found that the network structure can be classified into six types. In addition, Pierri et al. [12] focused on the fact that malicious topics such as fake news are increasing in social networks. They compared two networks, one with correct information and the other with false information. As a result, they found that there was a difference in the structure between the two networks.

In contrast, as previous research on SNSs, Watanabe et al. [13] conducted a network analysis on Twitter for two international cosmetic retail brands. They focused on the brand's tweets and obtained the brand's tweets (tweets containing the @brand account name) and tweets containing the brand's hashtags (tweets containing the brand name). Then, they created an "ego network" representing the information propagation of official brand tweets and a "hashtag network" representing tweets containing brand hashtags to identify the way information propagates. As a result of their analysis, they found that brands could send out messages; however they could not restrict communication on the network and might not control the spread of information. Zhao et al. [14] used social network analysis to identify highly influential hashtags and hashtag communities to restructure the fashion industry using the development of Internet technology. Specifically, they extracted hashtags from text data on SNSs related to specific events and visualized them as a network to identify hashtags with high junctions. Wang et al. [15] conducted a study on whether marketing activities on SNSs are effective in improving market sales. Specifically, they used price, brand, and relationship customer equity to examine whether marketing activities on SNSs lead to customer purchase. As a result of their analysis, they found that SNSs marketing activities contribute to improving customer equity. They also found that customer equity contributes to customer loyalty and future sales. In addition, Miyake et al. [16] identified community of consumers in SNSs and analyzed their consumer network. In particular, they targeted five competing fashion brands in the fashion market and clarify the difference in community structure. Han et al. [17] focused on the influence of social media and created a network of the top 10k influencers on Twitter. They demonstrated how to surface influencers related to specific fashion topics that companies would be interested in. In their study, they used the follow, retweet (spread function), and mention (replies) functions all to create the network.

By using the structure of the consumer network, it is possible to identify the influencers in that brand. These brand-specific influencers are commonly referred to as micro-influencers and are attracting a lot of attention. However, identifying micro-influencers in a brand does not determine whether the micro-influencers are influential in the whole industry. We named these industry-wide influencers as "mega-influencers." To identify the mega-influencers, it necessary to clarify the network structure using consumers in the entire industry. We speculate that micro-influencers and mega-influencers are not necessarily the same. In addition, few related studies have identified both micro-influencers and mega-influencers. In particular, research on SNSs targeting fashion and cosmetic brands has focused on hashtags, but not so much on users. In this study, we focused on both micro-influencers and mega-influencers.

II. Purpose

The purpose of this study is to identify the consumer network structure of several cosmetic brands and to compare the differences in network structure among brands. In addition, it is compared whether there is a difference between micro-influencers and mega-influencers. The data of specific cosmetic brand accounts were collected from Twitter (one of the SNSs). Following, a consumer network for each brand and compared the consumer network structure for each brand was created, also identified the micro-influencers for each brand. Subsequently, the consumer networks of each brand were used to combine followers to create a consumer network for the entire cosmetics industry. Then, we identify and compare the mega-influencers that influence the entire industry.

III. Dataset and Data Processing

In this study, consumer networks were constructed using data from Twitter, one of the SNSs. We used the Twitter API to obtain data. Then, we processed the data to create a network.

A. Hypothesis for Influencers on Twitter

In this study, we defined the network structure of arbitrary followers on Twitter, as shown in Fig. 2. The area in Fig. 2 represents the number of users in the cosmetics industry. First, we define general consumers as Twitter users. Among these general consumers, micro-influencers influence each brand. Similarly, there are mega-influencers in the industry as a whole among general consumers. Based on this hypothesis, we extracted data from Twitter API.

B. Dataset

The targeted official accounts of the six cosmetic brands are in Table I. The year of establishment in Table I is the establishment of a cosmetic line. The following three criteria were used to select target brands:

- Have a store in a department store.
- Have an official account on Twitter.
- Have more than 100,000 followers on Twitter.

The approximate number of followers and brand lineages of the six target brands are shown in Table II (as of January 8, 2022).

C. Data Processing

Next, we explain how to obtain data for each brand introduced in Section B. To clarify the network structure of consumers and influencers in this study, the following data are obtained as Fig. 3.

To explain the data structure, a brand A is picked up. First, we obtain a list of followers for a given brand. In Fig. 3, f_1 to f_n represent the followers of brand A. Next, the list of users who follow the followers of brand A is obtained. In Fig. 3, the followers of brand A's follower f_1 are from ff_1 to ff_m. Then, we created a network using the data of the followers and users who follow the followers.

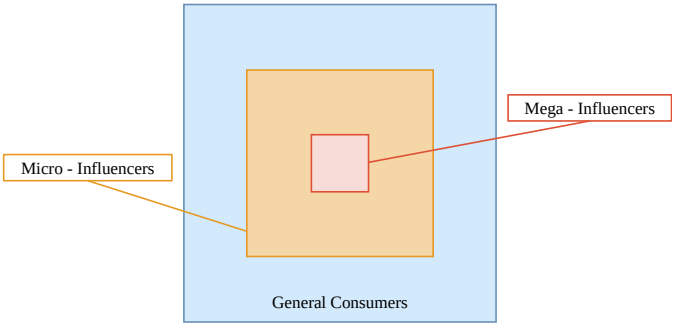


Fig. 2. Definition of Follower Hierarchy.

TABLE I. SUMMARY OF THE SIX BRANDS SELECTED

Brand Names	Twitter ID	Trademark	Founded
ADDICTION	@StyleADDICTON	KOSE	2009
Shu Uemura	@shuuemurajp	L'Oreal	1968
Paul & Joe	@PaulJoeBeauteJP	KOSE	2002
ETVOS	@etvos_jp	LVHM	2007
CLINIQUE	@CliniqueJp	ESTEE LAUDER.	1968
LANCOME	@Lancome_JP	L'Oreal.	1935

TABLE II. NUMBER OF FOLLOWERS OF THE TARGET BRANDS

No.	Brand Names	Followers (10 thousand)	Fashion Line
1	ADDICTION	12.1	Trendy
2	Shu Uemura	22.6	Trendy
3	Paul & Joe	11.4	Cute
4	ETVOS	10.2	Natural
5	CLINIQUE	11.5	Natural
6	LANCOME	16.2	Ellegance

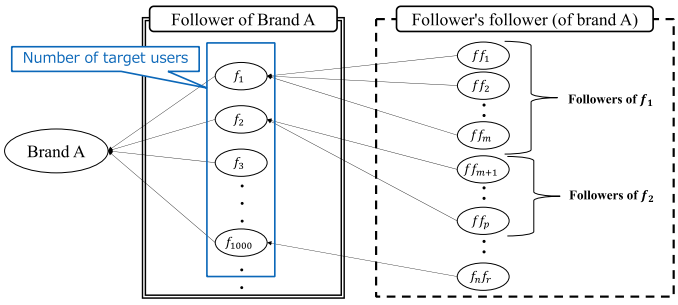


Fig. 3. Data Acquisition Flow.

In this study, it is difficult to obtain all of the users who follow Brand A's followers, and it is impossible to obtain information on accounts that are not public. Therefore, among the followers, we focus on the followers of 1,000 accounts that are public accounts (blue box in Fig. 3). We used the first 1,000 accounts that were followed in a short period of time, rather than randomly selecting accounts. In addition, we excluded sweepstakes-only accounts that include the words "sweepstakes, winning" in the description from the scope of this study. To create a consumer network on Twitter, we used the brand's followers (double-lined box in Fig. 3) as the nodes of the network among the acquired data. For the edges, we used users who followed the brand's followers (dotted boxes in Fig. 3).

Moreover, we weighted the edges in the following manner. As shown in Fig. 4, the followers of brand A are f_1 and f_2. From Fig. 4, the followers of f_1 are ff_1, ff_2, ff_3, and the followers of f_2 are ff_1, ff_3, ff_4. f_1 and f_2 have two common followers, ff_1 and ff_3. We used the number of common followers as the weights of the edges.

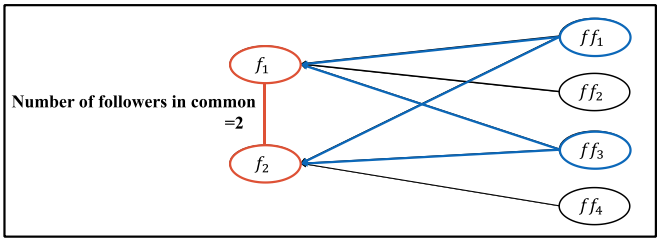


Fig. 4. Edge Weighting Method.

IV. VISUALIZATION OF FOLLOWER NETWORKS AND IDENTIFICATION OF MICRO-INFLUENCERS AND MEGA-INFLUENCERS

In this study, we first created a follower network for each brand and identified the network structure. In addition, we identified influential micro-influencers and communities within the brand. Next, we created a network for the entire cosmetic brand. For the overall network, we used the nodes that were ultimately used in each brand's network. For the structure of the network, we used the Fruchterman-Reingold model [18] which is a dynamical model. It is also used PageRank [19] as a metric to identify the influencers. Modularity [20] was used to identify the user community. For node reduction, we used the magnitude of the edge weights described in Section 3B.

A. Follower Network Structure and Influencers of Each Cosmetic Brand

In this study, we first visualized the follower network of each cosmetic brand. The number of nodes used is listed in Table III. To create the network, we used node reduction for accounts with more than one degree node. Edge weights were used to reduce the number of nodes. The number of nodes used in the network was set to approximately 30% of the initial number of nodes for each brand.

1) Network structure of "ADDICTION" and identifing of micro-influencers: Fig. 5 shows the visualization of the follower network of ADDICTION using all nodes with a degree greater than 1. From the degree distribution and Fig. 5, it was confirmed that the follower network of ADDICTION was scale-free. Fig. 6 shows a visualization of the addiction follower network after branch cutting. The color of the nodes indicates the community based on modularity. The size of the nodes is proportional to the size of the PageRank.

The average degree was 53.0. The number of communities detected by the modularity was three. For each community, Community 1 is brown, Community 2 is light blue, and Community 3 is beige. The density of the network was 0.226. The top five users with the highest PageRank are listed in Table IV. The highest degree in the ADDICTION network was 142 and the lowest degree was 12. From Table IV, it was confirmed that micro-influencers with high PageRank in the ADDICTION network also tend to have high degree.

TABLE III. NUMBER OF NODES USED FOR EACH BRAND

No.	Brand Names	Initial Number of Nodes	Final Number of Nodes
1	ADDICTION	767	236
2	Shu Uemura	753	222
3	Paul & Joe	709	210
4	ETVOS	761	229
5	CLINIQUE	809	241
6	LANCOME	791	237

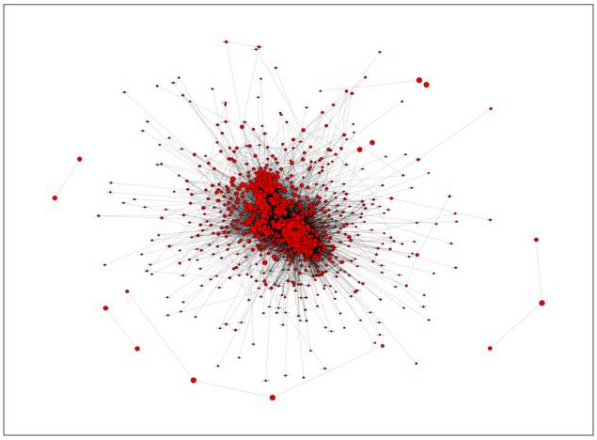


Fig. 5. Whole Consumer Network of ADDICTION.

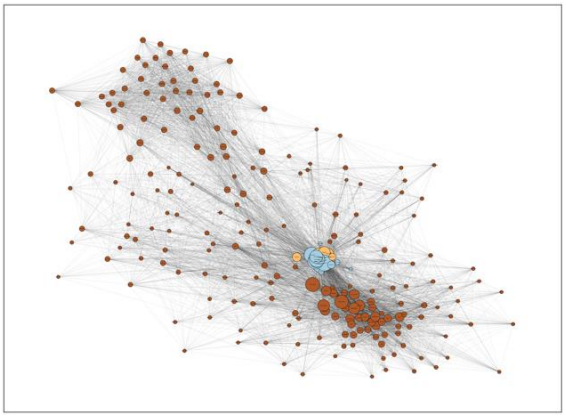


Fig. 6. ADDICTION Follower Network.

TABLE IV. TOP 5 USERS WITH HIGH PAGERANK (ADDICTION)

Rank	Users	PageRank	Degree	Community
1	A	0.0352	106	3
2	B	0.0316	111	2
3	C	0.0312	91	3
4	D	0.0296	109	2
5	E	0.0264	75	2

Based on the users' tweeting tendencies, each community is characterized as, community 1 as a community of cosmetic lovers, community 2 as a community of interested in beauty and cosmetics, and community 3 as a daily tweet community. In addition, the density of communities 2 and 3 is higher than that of community 1, and there is also a difference in the degree, so we determined that the network structure of the addiction is a "Centralized Network."

2) Network structure of "shu uemura" and identifing of micro-influencer: Fig. 7 shows the visualization of the follower network of Shu Uemura using all nodes with a degree greater than 1. From the degree distribution and Fig. 7, it was also confirmed that the follower network of Shu Uemura was scale-free.

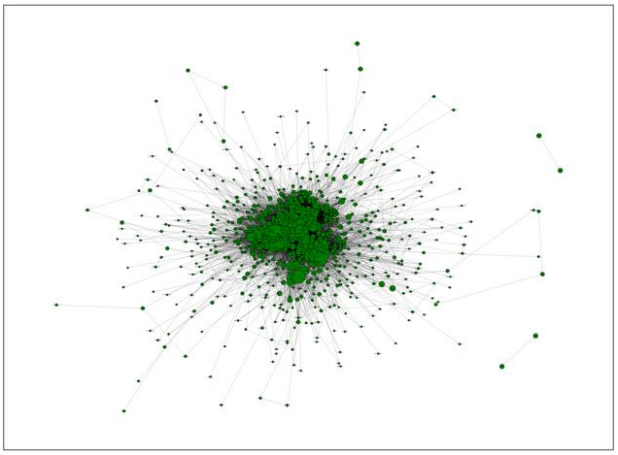


Fig. 7. Whole Consumer Network of Shu Uemura.

Fig. 8 visualizes the Shu Uemura follower network after branch cutting. The average degree was 49.1. The number of communities detected by the modularity was four. For each community, Community 1 is light purple, Community 2 is brown, Community 3 is pink, and Community 4 is light blue. The average network density was 0.222.

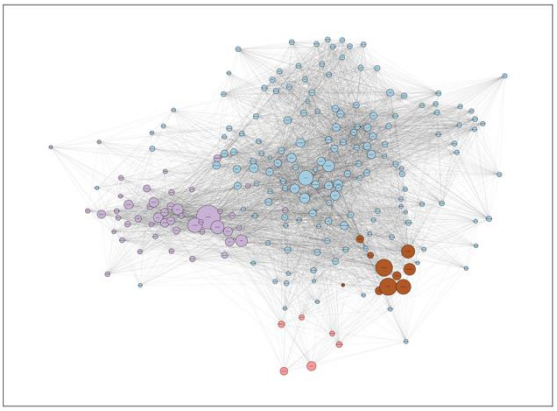


Fig. 8. Shu Uemura Follower Network.

TABLE V. TOP 5 USERS WITH HIGH PAGERANK (SHU UEMURA)

Rank	Users	PageRank	Degree	Community
1	F	0.0563	120	1
2	G	0.0299	49	2
3	H	0.0288	101	2
4	I	0.0223	22	2
5	J	0.0216	118	4

Table V lists the top five users with the highest PageRank. The highest degree in the Shu Uemura network was 132 and the lowest degree was 17. From Table V, it was confirmed that the degree of micro-influencers varied in the Shu Uemura network.

Based on the users' tweeting tendencies, we determined that community 1 is the cosplayer community (cosplayers are people who dress like anime and manga characters), community 2 is the homemaker community, community 3 is the community that likes cosmetics, and community 4 is the anime fan community. In addition, because the communities had many followers who were related to each other and the average degree was smaller than the others, we named them to be "Omnidirectional Network."

3) Network structure of "paul & joe" and identifing of micro-influencers: Fig. 9 shows the visualization of the follower network of Paul & Joe using all nodes with a degree greater than 1. From the degree distribution and Fig. 9, it was confirmed that the follower network of Paul & Joe was scale-free. Fig. 10 shows a visualization of the Paul & Joe follower network after branch cutting. The average degree was 48.9%. The number of communities detected by the modularity was four. For each community, Community 1 is light purple, Community 2 is brown, Community 3 is light blue, and Community 4 is pink. The average network density was 0.234.

Table VI lists the top five users with the highest PageRank. The highest degree in the Paul & Joe network was 134 and the lowest degree was 14. From Table VI, it was confirmed that the Paul & Joe network tends to have a high degree of micro-influencers, except for user N.

Based on the users' tweet tendencies, we determined that community 1 was a daily tweet community, community 2 was the cosmetic lovers' community, community 3 was the color analysis community, and community 4 was communities derived from communities 1, 2, and 3. As shown in Fig. 7, communities 1 and 4 were formed by deriving from a larger community consisting of communities 2 and 3. In addition, the average degree is not high. We declared Paul & Joe network to be the "Centralized Network."

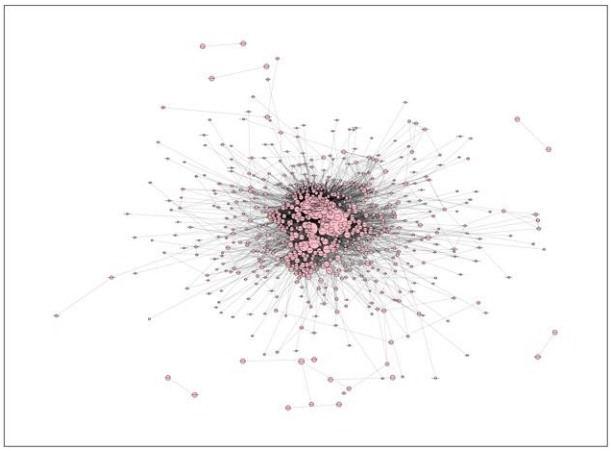


Fig. 9. Whole Consumer Network of Paul & Joe.

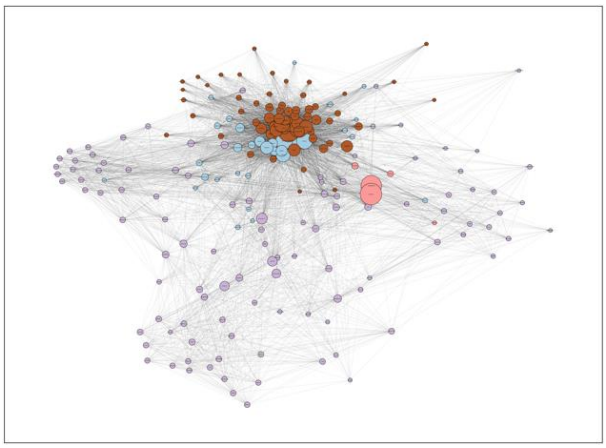


Fig. 10. Paul & Joe Follower Network.

TABLE VI. TOP 5 USERS WITH HIGH PAGERANK (PAUL & JOE)

Rank	Users	PageRank	Degree	Community
1	K	0.0390	106	4
2	L	0.0387	126	2
3	M	0.0366	119	2
4	N	0.0353	54	4
5	O	0.0290	105	2

4) Network structure of "ETVOS" and identifing of micro-influencers: Fig. 11 shows the visualization of the follower network of the ETVOS using all nodes with a degree greater than 1. From the degree distribution and Fig. 11, it was confirmed that the follower network of the ETVOS was scale-free. Fig. 12 shows a visualization of the ETVOS follower network after branch cutting. The average degree was 51.1. The number of communities detected by the modularity was five. For each community, Community 1 is light purple, Community 2 is light blue, Community 3 is pink, Community 4 is brown, and Community 5 is brown. The average density of the ETVOS network is 0.224.

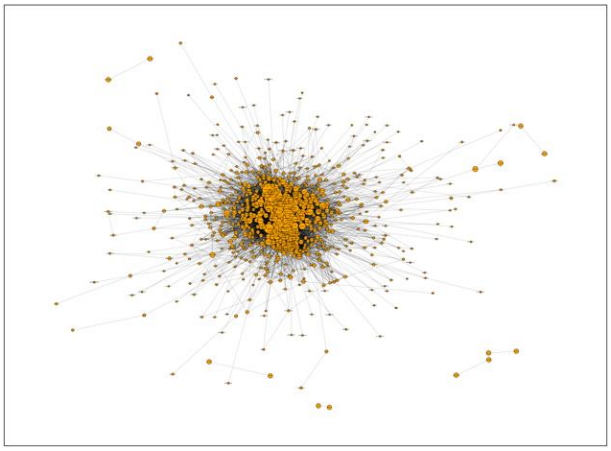


Fig. 11. Whole Consumer Network of ETVOS.

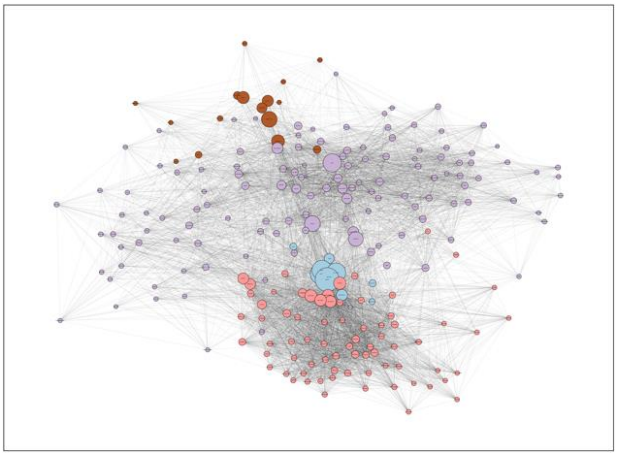


Fig. 12. ETVOS Follower Network.

TABLE VII. TOP 5 USERS WITH HIGH PAGERANK (ETVOS)

Rank	Users	PageRank	Degree	Community
1	P	0.0459	115	2
2	Q	0.0455	99	2
3	R	0.0383	76	2
4	S	0.0314	112	2
5	T	0.0296	38	2

Table VII lists the top five users with the highest PageRank. The highest degree in the ETVOS network was 170 and the lowest degree was 12. From Table VII, it was confirmed that the degree of micro-influencers varied in the ETVOS network. In particular, user T had a low degree but a high PageRank. In addition, we found that all the micro-influencers belonged to the same community. Based on users' tweet tendencies, community 1 was a daily tweet community, community 2 was a homemaker community, community 3 was an imprisoned community, community 4 was a cosmetic lovers' community, and community 5 was a lot of interest community. As the density of the ETVOS network tended to be low compared to the average degree of the ETVOS network, and as Fig. 8 shows, there was a connection between the communities. We named the network structure of ETVOS the "Omnidirectional Network."

5) Network structure of "CLINIQUE" and identifing of micro-influencers: Fig. 13 shows the visualization of the follower network of CLINIQUE using all nodes with a degree greater than 1.

From the degree distribution and Fig. 13, the follower network of CLINIQUE was scale-free. Fig. 14 shows a visualization of the CLINIQUE follower network after branch cutting. The average degree was 68.3. The number of communities detected by the modularity was three. For each community, Community 1 is light blue, Community 2 is beige, and Community 3 is brown. The average density of the CLINIQUE network is 0.284.

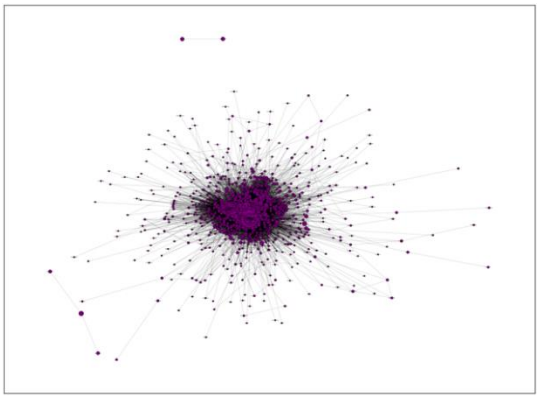


Fig. 13. Whole Consumer Network of CLINIQUE.

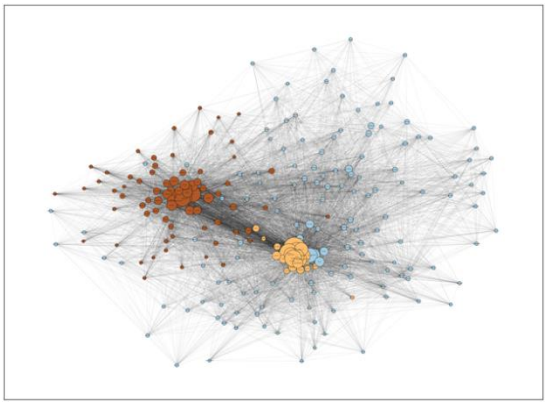


Fig. 14. CLINIQUE Follower Network.

TABLE VIII. TOP 5 USERS WITH HIGH PAGERANK (CLINIQUE)

Rank	Users	PageRank	Degree	Community
1	U	0.0508	158	2
2	V	0.0377	131	2
3	W	0.0300	123	2
4	X	0.0289	98	2
5	Y	0.0254	126	3

Table VIII lists the top five users with the highest PageRank. The highest degree in the CLINIQUE network was 185 and the lowest degree was 22. From Table VIII, it was confirmed that micro-influencers of the CLINIQUE network tended to be high degree, although their PageRank varied. Based on users' tweet tendencies, we determined that community 1 had many hobbies community, community 2 was a homemaker community, and community 3 was a cosmetic lovers' community. From Fig. 9, it seems that communities 1, 2, and 3 are divided into two groups. We determined that the connections between users in these two communities were high, based on the average degree of the network. Therefore, we named CLINIQUE network to be "Dual Network."

6) Network structure of "LANCOME" and identifing of micro-influencers: Fig. 15 shows the visualization of the follower network of the LANCOME using all nodes with a degree greater than 1. From the degree distribution and Fig. 15, the follower network of the LANCOME was scale-free. Fig. 16 shows a visualization of the LANCOME follower network after branch cutting. The average degree was 58.0. The number of communities detected by the modularity was four. For each community, Community 1 is brown, Community 2 is pink, Community 3 is light blue, and Community 4 is light purple. The average density of the LANCOME network is 0.246.

Table IX lists the top five users with the highest PageRank. The highest degree in the LANCOME network was 165 and the lowest degree was 22. From Table IX, it was confirmed that the degree was proportionally high to the PageRank of micro-influencers in the LANCOME network. Based on users' tweet tendencies, we determined community 1 as a cosmetic lovers' community, community 2 had a lot of hobbies community, community 3 was a cosmetic lovers and sweepstakes community, and community 4 was a talent lovers' community. From Fig. 16, we found that there are two large communities, one by Community 1, 2, and 3 and the other by Community 4.

Although the average degree of the network is not high, there are connections between the two large communities. Therefore, we named the network structure of LANCOME as a "Dual Network."

B. Visualization of Follower Networks across the Cosmetics Industry and Identify Mega-influencers

Next, in order to find mega-influencers, we created a follower network for all six brands using the follower users used in Section A. We reduced the number of nodes by using the edge weights calculated in Section 3-B. We used 357 nodes for the final visualization of the 1,277 nodes remaining in Section A. Table X shows the breakdown of the number of nodes for each brand used in the final network visualization.

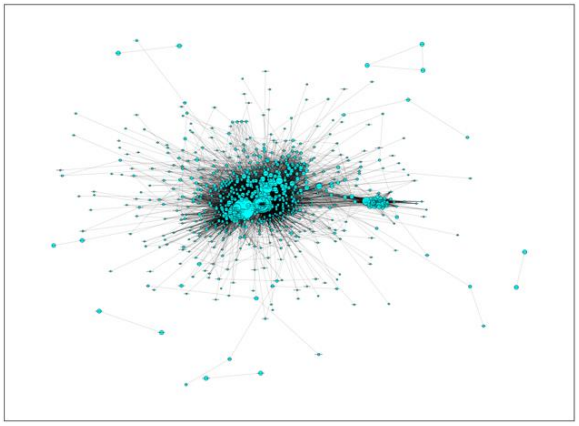


Fig. 15. Whole Consumer Network of LANCOME

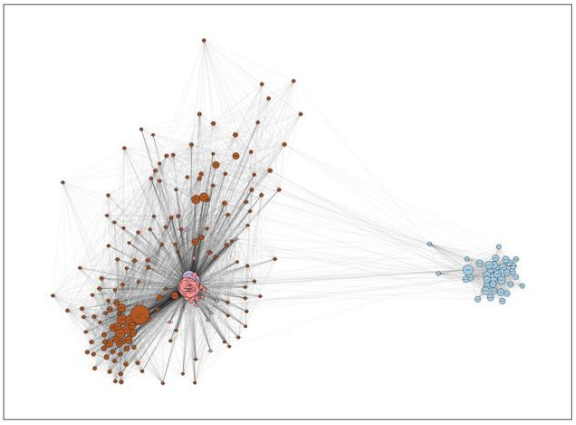


Fig. 16. LANCOME Follower Network.

TABLE IX. TOP 5 USERS WITH HIGH PAGERANK (LANCOME)

Rank	Users	PageRank	Degree	Community
1	Z	0.0330	125	1
2	AA	0.0265	115	2
3	AB	0.0234	165	2
4	AC	0.0232	101	1
5	AD	0.0209	132	2

TABLE X. NODE COLOR AND NUMBER OF NODES FOR EACH BRAND

Brand Names	Colors	Number of Nodes
ADDICTION	Pink	47
Shu Uemura	Yellow	39
Paul & Joe	Light green	42
ETVOS	Light blue	36
CLINIQUE	Light purple	58
LANCOME	Orange	106
Multiple	Brown	29

Fig. 17 shows the results of visualizing the network using the nodes listed in Table X. Some users followed more than one brand in Fig. 17 (29 nodes as shown in Table X). The network average degree was 47.3.

As a result of using modularity for the entire brand network, we detected five communities (Fig. 18). Community 1, light blue; Community 2, light green; Community 3, pink; Community 4, orange; Community 5, purple; and Community 6, brown. From Fig. 17 and 18, we found that the upper-right community (community 6) follows LANCOME. In addition, we found that the lower left community (community 3) followed the Shu Uemura. So, there are communities in the whole network where only the followers of a particular brand stick together.

Table XI lists the top ten users with the highest PageRank. "Brand" in Table XI represents the brands that each user is following. Comparing Table XI with the micro-influencers for each brand, it was confirmed that half of the mega-influencers were not included in the micro-influencers for each brand. Comparing the percentage of communities, we found that the percentage of communities in Community 1 (the orange community in Fig. 18) was high, indicating that it was the central community in the overall network.

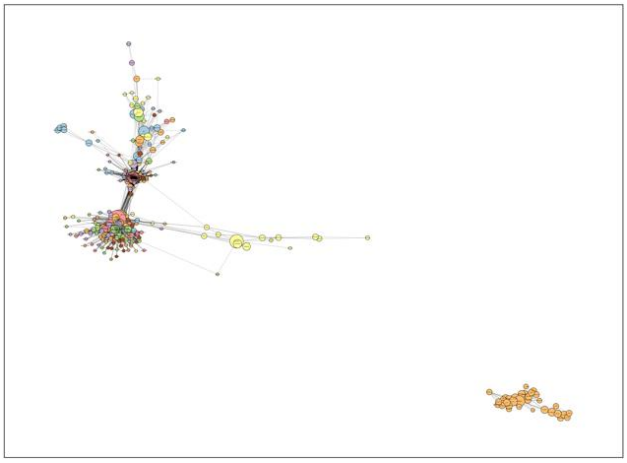


Fig. 17. Network of Coloration by Brand.

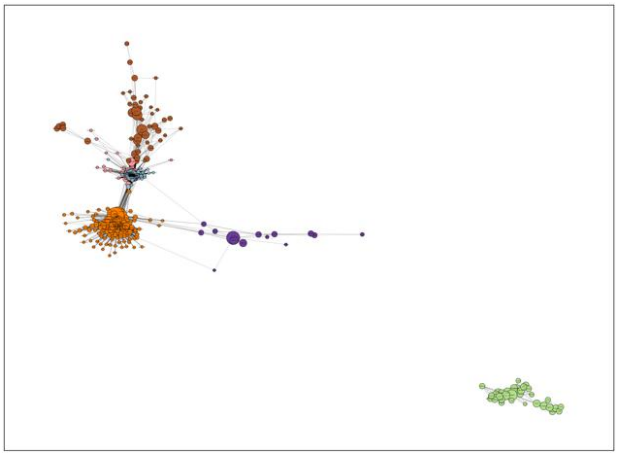


Fig. 18. Network Diagram of the Entire Brand.

TABLE XI. TOP 10 USERS WITH HIGH PAGERANK

Rank	Users	PageRank	Community	Brand
1	Z	0.0227	1	LANCOME
2	AE	0.0177	1	ADDICTION
3	AF	0.0129	1	CLINIQUE
4	Y	0.0125	1	CLINIQUE
5	AG	0.0114	1	ETVOS
6	F	0.0107	3	Shu Uemura
7	AH	0.0104	1	LANCOME
8	U	0.0104	5	CLINIQUE
9	AI	0.0102	1	CLINIQUE
10	A	0.0090	2	ADDICTION

V. DISCUSSION

First, it is confirmed that all brand consumer networks were scale-free networks, as shown in Fig. 5, 7, 9, 11, 13, and 15. Therefore, we considered that some micro-influencers mainly spread the information to other users who follow each micro-influencer.

Fig. 19 summarizes the results of classifying the follower network of the six brands into three network structures. The "Omnidirectional Network" refers to a network in which the entire community is universally connected. The "Dual Network" is a network in which the entire network is divided into two major communities. In these network structures, it is obtained those characteristics similar to those of the small-world in the degree distribution after node reduction. The "Centralized Network" is a network in which nodes are distributed to derive from one central community. From Tables V and VII, we confirmed that the micro-influencers of Shu Uemura and ETVOS, which are "Omnidirectional Network," varied degree. On the other hand, we confirmed that degrees of the micro-influencers of the four brands with "Dual Network" and "Centralized Network" tended to be high in proportion to PageRank from Tables IV, VI, VIII, and IX. In the omnidirectional network, there are various communities, and the communities are all related with each other. For this reason, we speculated that information tends to be transmitted easily even if the degree is not high, and PageRank may become high. On the other hand, "Dual Network" and "Centralized Network" are locally divided into one or two communities. Therefore, we considered that the value of PageRank was proportional to the value of degree.

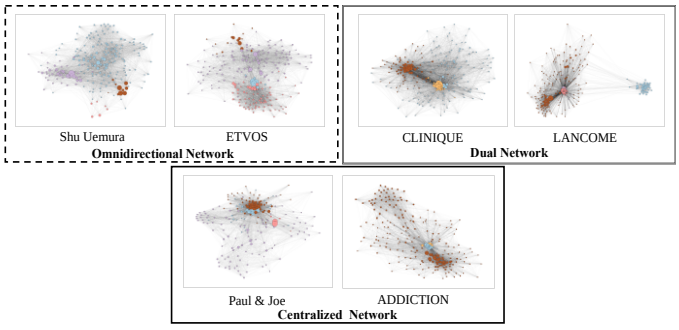


Fig. 19. Classifying Network Structures.

From the results of this classification and the comparison of the brand lineages in Table II, we inferred that the brand lines were not proportional to the network structure. As for the communities of each brand, it is observed a community of tweets about sweepstake entries in all networks, which was not directly related to cosmetics. These communities were unique to Twitter and considered the effect of a campaign to increase followers and product recognition. As a characteristic community, Shu Uemura had a community that liked animation and comic, and LANCOME had a community that liked celebrities. Shu Uemura sells cosmetics with tie-ups with manga and anime characters every winter, and LANCOME has Japanese celebrities as ambassadors, actively appearing in commercials and advertisements. Therefore, we consider that there is a good possibility that users, who prefer the tie-up side, but not the cosmetics themselves, are interested in the products and become fans. In contrast, Paul & Joe and ADDICTION specialize in communities of users who are interested in cosmetics and beauty. Paul & Joe's cosmetics with cat motifs are popular, and ADDICTION's single eyeshadow colors are abundant and popular. Therefore, we considered that they might be more likely to attract collectors than other brands.

Based on the tweeting tendencies of users belonging to each community, we roughly classified them into communities with similarities by community detection. However, some of the derived communities were inconsistent in their tweet content and self-introductions. It is assumed that this was due to the modularity of the community detection.

Table XII summarizes the results of the micro-influencers for each brand. As shown in Fig. 17, some users followed more than one brand. However, Table XII shows that there were no overlapping users among the six brands. Therefore, we inferred that user who are influential on one brand do not necessarily influence other brands. In addition, comparing Tables XI and XII, which show the mega-influencers, only five users (A, F, U, Y, and Z) were included in both tables. Therefore, we inferred mega-influencers were not necessarily micro-influencers for each brand. In contrast, Table XI shows that all mega-influencers followed only one brand. Given these facts, we considered whether a person following multiple brands was not a criterion for being a mega-influencer.

TABLE XII. TOP 5 MICRO-INFLUENCERS FOR EACH BRAND

	ADDICTION	**Shu Uemura**	**Paul&Joe**	**ETVOS**	**CLINIQUE**	**LANCOME**
1	A	F	K	P	U	Z
2	B	G	L	Q	V	AA
3	C	H	M	R	W	AB
4	D	I	N	S	X	AC
5	E	J	O	T	Y	AD

Furthermore, none of the users following Paul & Joe are included in Table XI. From the node breakdown in Table X, the number of nodes was not extremely low, and the density of brand consumer networks was not extremely small. One of the reasons for this was that the users who followed Paul & Joe were biased toward accounts that specialized in cosmetics. Contrarily, although Shu Uemura had the smallest number of nodes in the overall network, users who followed Shu Uemura were included in the top ten mega-influencers. Considering the fact that Shu Uemura's followers come from various lines of users, we considered that even in the cosmetics industry, users who were connected to various lines of users were more likely to become mega influencers who influence the whole network. However, as shown in section 4-B, even in the entire brand network, there were communities that consisted only of followers of a particular brand. We thought that users who belonged to such communities were not necessarily influential to the entire industry. Therefore, to identify mega-influencers, we considered it necessary to take into account their community affiliation even if their PageRank was high.

VI. CONCLUSION AND FUTURE WORK

In this study, we identified and compared the consumer network structures of six cosmetic brands. We also identified and compare the micro-influencers for each brand with the mega-influencers for the brands as a whole. We used network analysis. The consumer network was visualized using the Fruchterman-Reingold model, PageRank, and modularity to reveal the community.

As a result of the analysis, we classified the six cosmetic brands into three network structures. We named the three networks "Omnidirectional Network," "Centralized Network," and "Dual Network." It was also found that the micro-influencers with high PageRank and high degree for each brand were different. After visualizing the entire network, we classified the network into six communities. Of the six brands, only Paul & Joe's followers were not included in the top 10 mega-influencers with high PageRank. Comparing micro-influencers and mega-influencers, it was found that micro-influencers are not always mega-influencers.

In future work, we believe that a generalization of consumer networks in the cosmetics industry will be possible by conducting the same analysis for cosmetic brands that match the same conditions as in this study and comparing the results. In particular, the recent outbreak of COVID-19 has led to a decline in sales in the cosmetics industry in Japan, so we consider this to be an effective marketing strategy. In addition, although we used modularity for community detection in this study, some communities could not be judged well based on users' tweeting tendencies. Therefore, we believe that network analysis using metrics other than modularity will allow us to evaluate networks from different perspectives.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 19K01945, 21H04600 and 21K13385.

REFERENCES

[1] ICT Research & Consulting Inc., Survey on SNSs Usage Trends in 2020 , https://ictr.co.jp/report/20200729.html/, last viewed on Jan. 29, 2022.

[2] Statista, Most Popular Social Networks Worldwide as of October 2021, Ranked by Number of Active Users, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, last viewed on Jan. 20, 2022.

[3] Ministry of Internal Affairs and Communications of Japan, Results of Telecommunications Usage Trends Survey in 2018, https://www.soumu.go.jp/johotsusintokei/statistics/data/190531_1.pdf, last viewed on Jan. 29, 2022.

[4] Carlos H.G. Ferreira, F. Murai, Ana P. C. Silva, J. M. Almeida, M. Trevisan, L. Vassio, M. Mellia and I. Drago, "On the Dynamics of Political Discussions on Instagram: A Network Perspective," Online Social Networks and Media, Vol. 25, 2021.

[5] X. W. Wang, Y. M. Cao, C. Park, "The Relationships Among Community Experience, Community Commitment, Brand Attitude, and Purchase Intention in Social Media," International Journal of Information Management, Vol. 49, pp. 475-488, 2019.

[6] R. Huang, S. Ha, S. H. Kim, "Narrative Persuasion in Social Media: An Empirical Study of Luxury Brand Advertising," Journal of Research in Interactive Marketing, Vol. 12, pp. 274-292, 2018.

[7] E. J. Seo, J. W. Park, "A Study on the Effects of Social Media Marketing Activities on Brand Equity and Customer Response in the Airline Industry," Journal of Air Transport Management, Vol. 66, pp. 36-41, 2018.

[8] M. Girvan, M. E. J. Newman , "Community Structure in Social and Biological Networks," Proceedings of the National Academy of Sciences, Vol. 99 (12), pp.7821-7826, 2002.

[9] P. Chunaev, "Community Detection in Node-attributed Social Networks: A Survey," Computer Science Review, Vol. 37, 2020.

[10] M. Mizuno, H. Onishi, S. Shibuya, H. Yamamoto, "C2C Interactions in the Digital Media Environment : A Review and Perspective," Journal of Marketing Science, Vol. 26, pp. 7-39, 2018 (Japanese).

[11] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman and C. Espina, "Classifying Twitter Topic-Networks Uning Social Network Analysis," Social Media + Society, Vol. 3, 2017.

[12] F. Pierri, C. Piccardi, S. Ceri, "Topology Comparison of Twitter Diffusion Networks Effectively Reveals Misleading Information," Scientific Reports, Vol. 10, 2020.

[13] N. M. Watanabe, J. Kim, J. Park, "Social Network Analysis and Domestic and International Retailers: An Investigation of Social Media Networks of Cosmetic Brands," Journal of Retailing and Consumer Services, Vol.58, 2021.

[14] L. Zhao, C. Min, "The Rise of Fashion Informatics: A Case of Data-Mining-Based Social Network Analysis in Fashion," Clothing and Textiles Research Journal, Vol. 37 (2), pp. 1-16, 2019.

[15] H. Wang, E. Ko, A. Woodside, J. Yu, "SNS Marketing Activities as a Sustainable Competitive Advantage and Traditional Market Equity," Journal of Business Research, Vol. 130, pp373-383, 2021.

[16] S. Miyake, K. Otake, T. Namatame, "Analysis of Consumer Community Structure within Social Media -A Case Study of Competing Brands in Japanese Fashion Market-," International Academy of Business and Economics, Vol. 19, pp. 65-80, 2019.

[17] J. Han, Q. Chen, X. Jin, W. Xu, W. Yang, S. Kumar, L. Zhao, H. Sundaram and R. Kumar, "FITNet: Identifying Fashion Influencers on Twitter," Proceedings of the ACM on Human-Computer Interaction, Vol. 5, pp. 1-20, 2021.

[18] T. M. J. Fruchterman, E. M. Reingold, "Graph by Force-directed Placement," Software Practice and Experience, Vol. 21 (11), pp. 1129-1164, 1991.

[19] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bring Order to the Web," Technical Report of the Stanford Digital Library Technologies Project, 1998.

[20] M. E. J. Newman,"Modularity and Community Structure in the Networks," Proceedings of the National Academy of Science of the United States of America, Vol. 103, pp. 8577-8582, 2006.

Combining Multiple Seismic Attributes using Convolutional Neural Networks

Abrar Alotaibi[1], Mai Fadel[2], Amani Jamal[3], Ghadah Aldabbagh[4]
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia[1, 2, 3, 4]
Computer Science Department, College of Computer Science and Information Technology
Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia[1]
Department of Mechanical Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA[4]

***Abstract*—Seismic exploration involves estimating the properties of the Earth's subsurface from reflected seismic waves then visualizing the resulting seismic data and its attributes. These data and derived seismic attributes provide complementary information and reduce the amount of time and effort for the geoscientist. Multiple conventional methods to combine various seismic attributes exist, but the number of attributes is always limited, and the quality of the resulting image varies. This paper proposes a method that can be used to overcome these limitations. In this paper, we propose using Deep Learning-based image fusion models to combine seismic attributes. By using convolutional neural network (CNN) capabilities in feature extraction, the resulting image quality is better than that obtained with conventional methods. This work implemented two models and conducted a number of experiments using them. Several techniques have been used to evaluate the results, such as visual inspection, and using image fusion metrics. The experiments show that the Image-fusion Framework, using the Image Fusion Framework Based on CNN (IFCNN) approach, outperformed all other models in both quantitative and visual analysis. Its $Q_{AB/F}$ and *MS-SSIM* scores are 50% and 10%, respectively, higher than all other models. Also, IFCNN was evaluated against the current state-of-the-art solution, Octree, in a comparative study. IFCNN overcomes the limitation of the Octree method and succeeds in combining nine seismic attributes with a better-combining quality, with $Q_{AB/F}$ and $N_{AB/F}$ scores being 40% higher.**

***Keywords*—CNNs; neural networks; seismic attributes; seismic images; image fusion**

I. Introduction

Seismic data is a major source of information for Earth subsurface exploration and visualization. To gather seismic data, seismic waves are sent into the Earth's subsurface, and the resulting reflection is recorded. Using these reflections, the underlaying structural information is obtained and the earth subsurface can be modeled and visualized [1]. The data that are obtained from the seismic data, to supplement and enhance the geological/geophysical information, are referred to as seismic attributes [2]. They help to make the process of visualization more informative.

The current process used by engineers, archeologists, geologists, and other scientific scholars to develop accurate representations of the Earth's subsurface involves looking at the seismic images and their related seismic attributes, which is followed by the interpretation of huge volumes of data. The process is, however, bulky and makes it difficult to combine the various views into one comprehensive view that can efficiently exploit all the data included in each individual view and reduce the time taken in the process.

Various scholars have made major contributions to address the challenge of combining seismic attributes, including Octree, principal component analysis (PCA), cross-plotting, and volume blending [3], [4]. The most recent work by Al-Dossari et al. [4], show how the Octree color quantization algorithm can be extended to enhance the combined seismic attributes. However, the method has some limitations. For instance, the number of attributes is limited to a maximum of eight, the structural disposition of the attributes can affect the results, and the result of the combined image includes artifacts.

Image fusion can be described as the process of combining more than one input image that contains complementary information from related scenes, thus producing a composite image [5]. The input images are obtained from matching imaging devices, including various types of imaging devices, or from various other parameters such as infrared cameras and satellites. The resultant composite image is more useful in terms of the included information as compared to the individual images [6]. The techniques used in image fusion offer many benefits in different image processing tasks that rely on viewing more than one image of the same scene, such as object recognition and detection, as well as areas like digital photography and remote sensing, among others. Merging the key information of various input images into one fused image can be helpful in reducing the challenge of wasted time and enhancing the final results of the work [5]. The data enrichment offered by seismic attributes of seismic images is the same as in various other image fusion tasks, like remote sensing and medical imaging.

The recent development of deep learning (DL) has led to various experts in the field developing different image fusion techniques using the new technology. In this field, Machine Learning algorithms, afforded by deep learning, along with neural networks, are used to extract data and image representations. The use of Convolutional Neural Networks (CNN) is important in solving the conventional, manual method challenge of designing fusion techniques and choosing

activity-level metrics and fusion rules as it has the capability of learning features indirectly via data training. Because the tasks involved in image fusion are closely related to the classification challenges that CNNs excel in, they provide superior results [7], [8].

To create a DL method capable of combining any number of seismic attributes, this paper proposes using general image fusion models. The method involves extracting features of an image and then fusing them into a single image. It first obtains three-dimensional (3D) image information, with each piece of the three-dimensional information representing either the seismic attributes or the seismic (raw data) image. Based on this method, 3D data is sliced, and the resulting two-dimensional (2D) images are forwarded to the fusion model as inputs. The key data is then extracted from the input images by use of the convolutional layer to produce maps of the features. These maps are then fused to generate the output image and, lastly, the process outputs the data in form of a 3D image.

This paper includes experiments that compared the proposed technique by implementing two fusion models with other fusion models used previously by Alotaibi et al. [9], and then compares their results. It also compares the model's results against the results of Octree. The models used are a new kind of image fusion model developed to fuse all types of images and are not limited to any specific types of images. The reason for using pre-trained models is the lack of available datasets for seismic images with ground-truth fusion images, which hinders the training process. The pretraining helps solve the problem of training the CNN.

Our paper is structured as follows: In Section II, we briefly provide background information; in Section III, the proposed fusion method is introduced in detail; in Section IV, the experimental results are shown; in Section V the conclusion of our paper and discussion are presented.

II. BACKGROUND

A. Image Fusion Review

In its simplest terms, image fusion can be described as a technique used in image processing that involves merging more than one input image, obtained from multiple sensors, to produce a single superior image [5]. The process is used to reduce the volume of data, as well as to provide images that are more ideal and understandable by computers and humans. Image fusion also facilitates the collection of data from images derived from multiple sources to create high-quality fused images including all the spectral and spatial information [6].

The fused image must observe the following conditions: first, it must contain all of the relevant information; second, it must have clarity regarding every artifact and anomaly; and third, all errors and noise are eliminated. Some of the primary applications of image fusion are multi-focus image fusion, medical image fusion, and remote sensing image fusion [5].

The common approach used in image fusion includes acquiring multiple input images, registering the images, and then fusing them. The registration of the images includes detecting features, setting and comparing them, estimating the transformation models, and converting and re-sampling the image. The process includes fusion rules which are applied either as part of the image transformation models or as direct mathematical applications, such as choosing or averaging the maximum pixel value [10].

Image fusion can be classified into different categories according to the task(s) performed [11]. These are:

- Multi-exposure image fusion - combines images with various exposures to different lighting to produce superior images.
- Medical image fusion – combines images used in medical fields, like computed tomography (CT) and magnetic resonance imaging (MRI) to produce more informative images.
- Infrared/Visible light image fusion that combines images obtained using infrared radiation with visible light to produce images that are more informative.
- Multi-focus image - it combines images that include diverse focus depths to produce greater depth of visual field.

The deep learning-based image fusion approach has demonstrated huge potential in terms of enhancing the techniques used for image fusion due to the application of CNNs. The basic architecture of CNNs includes two main parts, the classifier and the feature extractor. The latter utilizes pooling and convolutional layers to obtain the relevant features of the inputs and signify them via activation maps, which support the step of image registration in the image fusion process. The former is used to execute fusion rules on the map, which supports the fusion part in the image fusing process. CNNs are also able to apply more than one fusion rule since they are trained on big datasets, thus avoiding one of the classical limitations of the fusion techniques. Fig. 1 presents the CNN's basic architecture.

B. General Image Fusion

Within the last few years, a new trend of image fusion research has emerged in which DL models are created to perform image fusion on all types of image fusion tasks. So far, two of such general image fusion models have been developed.

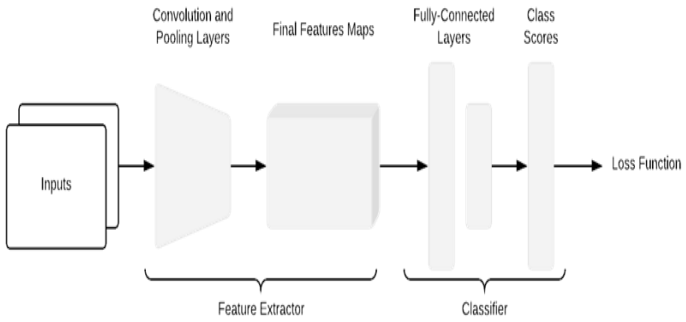


Fig. 1. Overview of CNNs Architecture.

One of these includes a Zero-learning image fusion model proposed by Lahoud and Susstrunk [12] to solve problems of image fusion using CNNs, such as the need to create a large dataset to train the network. The training process is time-consuming and expensive, requiring a lot of resources such as processing power, and a large amount of memory, energy, and time, which prevents the network from being able to perform other tasks. As a solution, the study proposes a novel image fusion model using earth CNNs. The pre-trained network is a network that has been trained on a large dataset and has saved the network's weights and biases to use on another task. Using a pre-trained network solves all the aforementioned problems by removing the need to create a dataset, saving time, and reducing cost. It also eliminates the training process to work on various image fusion tasks. The model operates as a two-scale decomposition image fusion model. Fig. 2, presents a schematic diagram of the proposed method. It follows specific steps, which include: (1) dividing the images into base and detail layers by applying a filter; (2) performing base layer fusion on saliency maps; (3) performing detail layer fusion on CNN feature maps using the Very Deep Convolutional Networks (VGG19) model trained with ImageNet [13]; and (4) fusing base and detail layers to acquire the final fused image.

The proposed model was tested against state-of-the-art models for medical image fusion, Infrared-Visible image fusion, and Multi-focus image fusion. The experimental results showed that the fusion model being developed was robust, and that it exceeded the current state-of-the-art image fusion models for specific tasks.

Zhang et al.[14] propose a general image-fusion framework using CNN (IFCNN) that takes full advantage of the convolutional layer capabilities as a feature extractor, as well as generating output images using a weighted average. The proposed framework is a novel solution to the problem of general-purpose image fusion to achieve state-of-the-art results with a fully convolutional neural network without the need for other techniques to complement it. The quality of the training dataset used in the model is far superior to other existing models, thus making the proposed framework a novelty among CNN models used for image fusion. The IFCNN contains three key modules, including the image reconstruction module, feature fusion module, and feature extraction module. Fig. 3 includes an illustration of the model architecture.

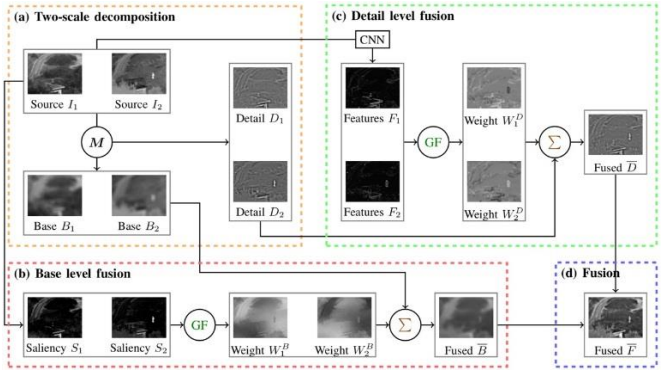


Fig. 2. Schematic Diagram of the Method [12].

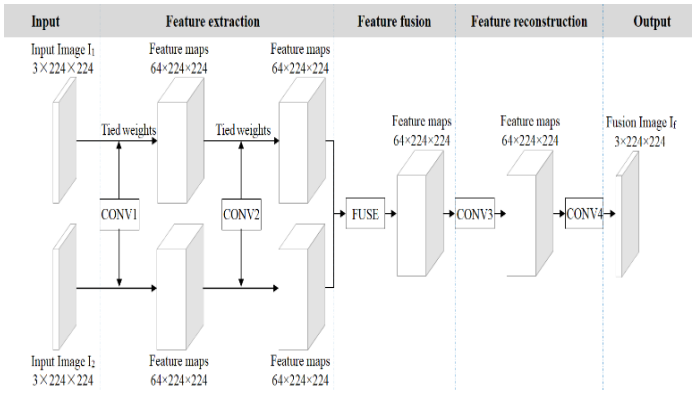


Fig. 3. IFCNN Architecture [14].

The model was also tested against state-of-the-art models used in multi-focus image fusion, multi-exposure image fusion, infrared-visible image fusion, and medical image fusion, and outperformed all of them.

Alotaibi et al. [9] proposed the use of DL models to combine seismic attributes. The work primarily investigated the performance of current state-of-the-art models, pre-trained for specific types of image fusion in combining three seismic attributes. The results showed that DeepFuse [15] has successfully combined three seismic attributes. DeepFuse is a model used for multi-exposure fusion. Prabhakar et al. [15] proposed DeepFuse, a novel model for multi-exposure fusion that takes an unsupervised approach in the fusion process. The authors also created and trained the network on a new benchmark dataset, improving the model's learning ability.

III. Proposed Technique

The image fusion technique proposed here helps to support the seismic data-merging and multiple seismic attributes along these lines: assume there are X inputs to the model, and X≥2, where X is three-dimensional images with similar sizes that are either seismic attributes or data, symbolized as I_{An} and I_R in that order, as $I_{An}|n\in\{1,2,3,\ldots,N\}$ as shown in Fig. 4. First, the I_{An} and I_R inputs are transferred to slicing functions to change the three-dimensional data (x,y,z) into two-dimensional data (x,y) with Z sum of images. The slicing function's outputs are sent to the fusion models as inputs, the model accept a group of images as inputs, including a single image from each I_{An} and I_R, beginning from z =1 up to Z. After every fused image is generated by the fusion models, and the calculation of the fusion metrics and image fusing is done, they are then converted into three-dimensional image information via the slicing function's reverse function.

A. Fusion Model

We will compare the performance of IFCNN and Zero-Learning models on the seismic image combining task. IFCNN is trained on a NYU-D2[16] dataset and Zero-Learning, using ImageNet weights for its layers, and pre-trained implementation to overcome the issue of a lack of labeled datasets where their ground truths are identified. This approach also benefits from the use of pre-trained models because it eliminates the need for model training, which reduces the time and resources required to implement the method. Additionally, using pre-trained models to combine

seismic attributes creates a method that is less complicated than all other existing methods since the existing methods require powerful workstations and high computational resources.

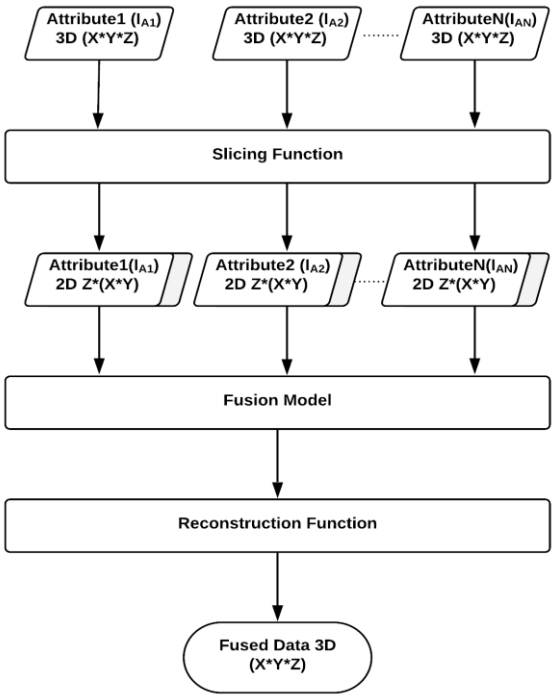


Fig. 4. Schematic Diagram of the Proposed Method.

B. Fusion Metrics

The metrics used to assess the performance of image fusion techniques proposed by Lahoud, Süsstrunk, et al. [12][14] compare the models' performance. Because the ground truth images are lacking, different non-referenced image fusion metrics are used in the evaluation of the performance of the models. Based on the study by Jagalingam and Hegde [17], the appropriate metrics to use are:

1) $N_{AB/F}$ (Modified Fusion Artifacts): for measuring artificial artifacts produced by the fusion.

2) EN (Entropy): for measuring the fused image's information content.

3) $Q_{AB/F}$ (Information Transfer): for measuring the overall information moved from source images to the fused one.

4) FMI (Feature Mutual Information): for measuring the dependencies existing between the fused image and input images.

5) MI (Mutual Information): measures the relationship of image intensity between the reference images and the fused images.

6) SSIM (Structural Similarity Index Measure): used for comparing the local patterns of pixel intensities between the reference images and the fused image.

7) MS-SSIM (Multi-scale structural similarity): used in measuring the expansion of the SSIM by merging luminance data at the highest resolution levels, with contrast and structural information at various down-sampled resolutions (scales).

IV. Experiments and Results

After conducting experiments on the models Zero-Learning and IFCNN, using pre-trained models published by Lahoud, Süsstrunk, et al. [12][14], the results were compared with those of DeepFuse and Octree [4][15]. Three key experiments are presented in this study. The first one compares the results from three models' results on combining three seismic attributes to find the best model, and the second one compares the results of Octree with the best fusion model.

Experiment 1 determined whether the proposed models are able to combine three different seismic images and then compared their results with DeepFuse. Experiment 2 showed the model's ability to combine up to nine attributes. Experiment 3 was used to compare the model results of combining eight seismic attributes against the results of Octree. To analyze the combined results, a visual comparison was done, along with a quantitative assessment, to check the visual representation characteristics, like color and quality, among other aspects of the fused image, together with the structural data.

A. Comparing Fusion Models

In the first experiment, a section from a marine block from the North Sea was used. The number of inputs X is 3; one of the inputs is a seismic image (I_R), and another is a skeletonization algorithm seismic attribute termed skeleton, produced by (I_{A1}) [18]; the third is a seismic attribute called coherence and is represented by (I_{A2}) [19]. The size of $I_{R,}$, I_{A1}, and I_{A2} is (876,221,271). The inputs and the combined results of the three models are presented in Fig. 5. The images are reduced and cropped to fit within the limits of the space available. The original images were used for the experiments.

To assess the success of the fusion of multiple seismic attributes, the fusion result should:

1) Identify unique events that appear in one attribute. In this experiment, the events are faults that appear in one of the inputs.

2) Preserve small details.

3) Reveal major common geo-bodies from all inputs. In our experiment, the geo-bodies are faults [1].

We will refer to each of these points as a Studied Property (SP).

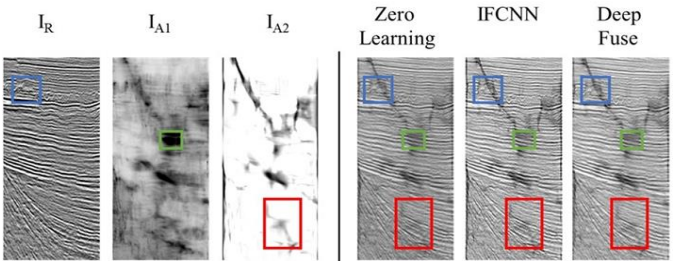


Fig. 5. Fusion Results after Combining Experiment 1 Results: The Left Side of the Line shows the Inputs while the Right Side Presents the Fusion Outputs.

For SP 1, described as Identifying unique events that appear in one attribute, as shown in Fig. 4 and marked in blue, Zero-Learning, IFCNN, and DeepFuse clearly present the events from all inputs. For SP 2, described as Preserving small details after the combining process, as shown in Fig. 4 marked in red, IFCNN maintained all of the details that other models could not keep. For SP 3, described as Revealing major common Geo-bodies in the fused image, Fig. 4 shows an example of a common geo-body, marked in green. All models displayed the geo-body but not all details are visible. Only IFCNN captured all of the details. IFCNN had the best performance, followed by Zero-Learning, which shows that general fusion models are capable of extracting and transferring important information from seismic images better than DeepFuse.

Experiment 1 fusion metrics results are presented in Table I. The figures in bold font represent the best performance results, while underlined figures indicate the second-best performance.

TABLE I. EXPERIMENTAL RESULTS

Fusion Metric	Model		
	Zero Learning	**IFCNN**	**DeepFuse**
EN↑	*7.185351*	**7.363646**	7.127503
MI↑	*21.55605*	**22.09094**	21.38251
$Q_{AB/F}$↑	0.35872	**0.55716**	*0.38529*
FMI↑	**0.840996**	0.831741	*0.840794*
SSIM↑	**0.4322**	0.413706	*0.428697*
MS-SSIM↑	0.785514	**0.870077**	*0.806129*
$N_{AB/F}$↓	*0.010318*	0.079795	**0.000752**

The Zero-Learning model's result included the top values for FMI, SSIM, and second-best values for EN, MI, and $N_{AB/F}$. Scoring high in EN, MI, and FMI, ≈0.84 values indicate that the fused images include huge volumes of data and that the model performs extremely well for extracting features. The model's EN and MI values are 2.4% less than the best model (IFCNN). Scoring high SSIM ≈ 0.43, and MS-SSIM ≈0.78 values, indicates that the fused images have maintained structural data and high image resolutions. The model's SSIM and MS-SSIM values are 1% and 9.7% respectively, less than the models with the highest values (IFCNN, DeepFuse). Scoring a high $Q_{AB/F}$, ≈ 0.35 value, shows that the fused image contains data transferred from the inputs. The model is ranked third for $Q_{AB/F}$ value, and less than the best model by 35.6%. Scoring a low $N_{AB/F}$ ≈ 0.01 value indicates less artificial fusion noise. The model is ranked second for $N_{AB/F}$ values and is more than the best model (DeepFuse) by 1×10^3 percent.

The IFCNN model's results included the top values in EN, MI, $Q_{AB/F}$, and MS-SSIM. The model's high MI and EN values indicate that the fused image includes some rich information. The model has a high FMI ≈ 0.83 value which is less than that of the model with best values (Zero-Learning) by 1%. The model has a high SSIM ≈ 0.41 and highest MS-SSIM ≈ 0.87 values; SSIM values are 4.2% less than the best model (Zero-Learning). The model has the highest $Q_{AB/F}$ ≈ 0.55 value and has a low $N_{AB/F}$≈ 0.079 value. It is ranked third for the $N_{AB/F}$ values, and the value is more than the best model (DeepFuse) by 1×10^5 percent. The fused images contain all of the structural information from the inputs, every edge is clear, the inputs' texture and color are available, and no perceptible fusion noise is present.

For the DeepFuse model, its results included the top values for $N_{AB/F}$, and the second-best ones for FMI, $Q_{AB/F}$, SSIM, and MS-SSIM. The model has high EN and MI values. The EN and MI values are 3.3% less than the model with the best EN and MI values (IFCNN) and the model's FMI ≈ 0.86, and are less than the model with best values (Zero-Learning) by 0.02%. The model has high SSIM ≈ 0.42 and high MS-SSIM ≈ 0.80 values. The SSIM and MS-SSIM values are 1% and 7.8% respectively, less than the models with the highest values (Zero-Learning and IFCNN). The model has a high $Q_{AB/F}$ ≈ 0.38 value but is less than the best model (IFCNN) by %30. The model has the lowest $N_{AB/F}$ ≈ 0.0004 value. The fused images contain all of the structural information from the inputs, every edge is clear, the texture and color from the inputs are presented clearly, and no perceptible fusion noise is present.

The Experiment showed that IFCNN is the best out of the investigated models, as it outperformed every other model and gave the best result for enhancing fault detection by combining seismic attributes. Zero-learning and DeepFuse were competing for second place and had comparable performance. We only considered IFCNN in the following experiments.

B. Combining more Attributes

The second experiment tested IFCNN's ability to combine multiple seismic attributes. We used a section from the Parihaka dataset and generated fusion results. In addition to seismic Detect and Skeleton attributes, six additional attributes were generated using an edge-preserving algorithm [20]. Fig. 5 presents the results of combining multiple attributes.

The fusion metrics of the results have been calculated and are presented in Table II.

TABLE II. FUSION METRICS' RESULTS FOR IFCNN FOR UP TO NINE INPUTS

Fusion Metric	Number of Inputs (Attributes)						
	3	**4**	**5**	**6**	**7**	**8**	**9**
EN↑	6.85	6.86	6.89	7.02	7.14	7.06	7.60
MI↑	20.9	27.4	34.48	42.13	50.00	56.55	60.87
$Q_{AB/F}$↑	0.56	0.48	0.36	0.32	0.31	0.25	0.21
FMI↑	0.86	0.84	0.84	0.82	0.82	0.82	0.76
SSIM↑	0.50	0.43	0.41	0.44	0.43	0.38	0.28
MS-SSIM↑	0.85	0.69	0.60	0.73	0.55	0.70	0.67
$N_{AB/F}$↓	0.05	0.05	0.01	*0.04*	0.13	0.08	0.17

The metrics values show that IFCNN had maintained good image quality for the fused images while increasing the number of attributes. First, the increase in EN and MI values with the increase of attributes shows that IFCNN combining results is rich in information. IFCNN maintained high FMI and MS-SSIM values while increasing the number of attributes, demonstrating that the fused images did not lose important information from individual inputs and that the fused images have a good structure. The $N_{AB/F}$ values increased while increasing the number of attributes, since increasing the number of inputs leads to increasing the amount of resulting fusion noise, but the values remained small. Finally, IFCNN $Q_{AB/F}$ and SSIM values decreased while increasing the number of attributes because, given that the fused image structural similarity to individual inputs and the information transfer rate from input to output will decrease with the increase of inputs, the decrease is to be expected.

Visually inspecting the combining results exhibit IFCNN's ability to increase the number of combined attributes without generating a large number of unwanted artifacts or diminishing visual information. After examining the combining results quantitatively using fusion metrics, and qualitatively by visual inspection, the combining results determined IFCNN's ability to successfully combine up to nine attributes, with the ability to combine more.

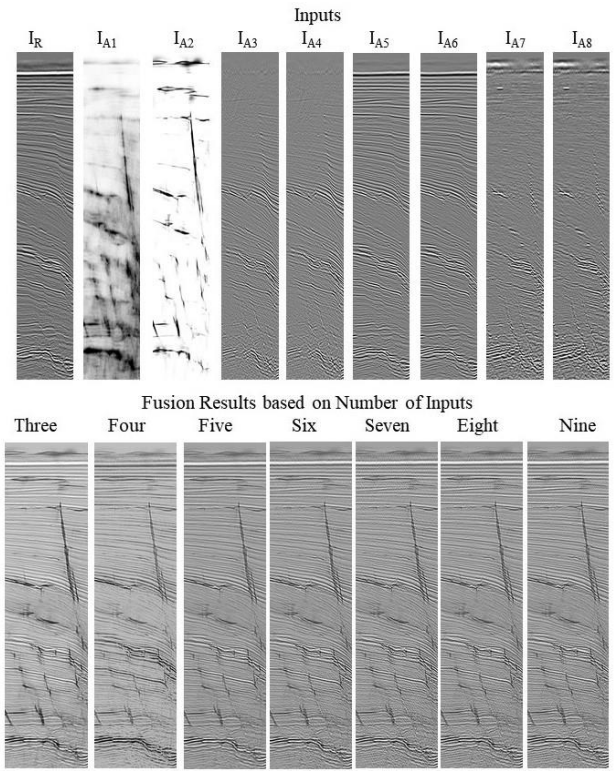


Fig. 6. A Sample of Combining Results of Multiple Attributes. Inputs: Seismic (IR), Detect Attribute (IA1), Skeleton Attribute (IA2), Sobel X Gradient Attribute (IA3), Prewitt X Gradient Attribute (IA4), Sobel Y Gradient Attribute (IA5), Prewitt Y Gradient Attribute (IA6), Sobel Z Gradient Attribute (IA7), Prewitt Z Gradient Attribute (IA8).

C. Comparing IFCNN and Octree

The third experiment was carried out to compare, in detail, the quality of combining results between IFCNN and Octree on fault detection. To fairly compare IFCNN and Octree, we used sections from a marine block and an F3 block to generate combining results of three and eight attributes respectively, Fig. 6 shows the input and output of both IFCNN and Octree for combining three attributes.

As shown in Fig. 6, IFCNN preserved more structural information in the resulting image than Octree, its combined image had less noise, and its results are more suited for the fault detection task. Then, we generated combining results for eight attributes and compared the performance of the two methods as shown in Fig. 8.

From Fig. 7, it can be seen that IFCNN's structural details are more vivid and that faults are more easily detectable since IFCNN uses a large number of filters to extract important features from individual attributes before combining. This helped IFCNN maintain high structural information with the increased number of attributes. The quality of the combining results can be quantified using the fusion metrics' values in Table III (better results are in bold).

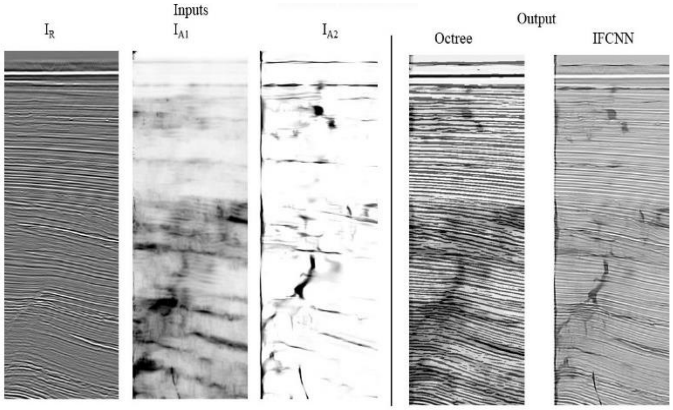


Fig. 7. Inputs: Seismic (IR), Detect Attributes (IA1), and Skeleton Attribute (IA2). Output: Combining Results of IFCNN and Octree.

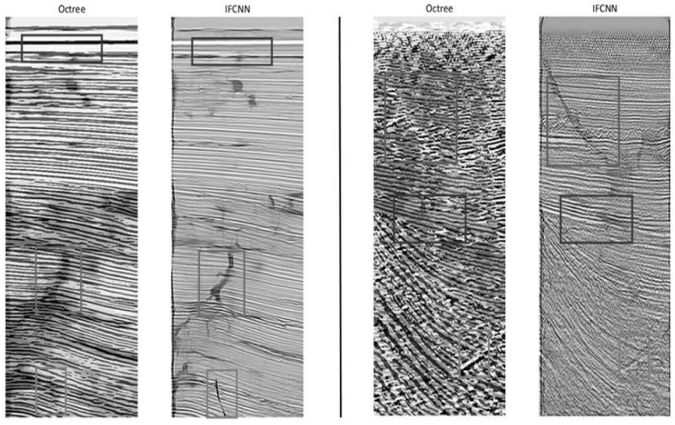


Fig. 8. Left Side: Combining Results of Three Attributes. Right Side: Combining Results of Eight Attributes.

TABLE III. FUSION METRICS' RESULTS FOR IFCNN AND OCTREE

Fusion Metric	3 Attributes		8 Attributes	
	IFCNN	**Octree**	**IFCNN**	**Octree**
EN↑	7.180216	**7.51156**	7.60924	**7.93288**
MI↑	21.54065	**22.5347**	60.8739	**63.4630**
$Q_{AB/F}$↑	**0.583826**	0.27468	**0.21114**	0.11362
FMI↑	0.850726	**0.87913**	**0.76648**	0.72074
SSIM↑	**0.457925**	0.25939	**0.28056**	0.18298
MS-SSIM↑	**0.878803**	0.41374	**0.67677**	0.26245
$N_{AB/F}$↓	**0.056377**	0.35625	**0.17012**	0.29739

The higher EN and MI values of the Octree method can be attributed to the color quantization algorithm, because Octree quantizes and inputs pixel values, which guarantees that all information from all inputs is condensed in the output, creating an information-rich image with a high amount of information. The FMI values are also affected by the color quantization algorithm, as observed in Table III. The Octree method had a higher value of FMI with a small number of inputs because the total amount of information can be quantized (condensed) without losing input information, but when quantizing a large number of inputs, some information from individual inputs is lost. Because of feature extraction, IFCNN managed to maintain high values of FMI with the increase of the number of inputs, because the important information is extracted from the images, and the fused image did not lose important information from individual inputs. Also, for the other metrics, it can be seen that the Octree method does not maintain high structural similarities between inputs and output, which is indicated by the low values of SSIM and MS-SSIM. Fig. 7 shows that the Skeleton attribute structure is not clearly visible in the Octree results. IFCNN, SSIM, and MS-SSIM values are 50% more than Octree, and this can be observed by IFCNN results maintaining a similar structure as input images. Also, IFCNN, $Q_{AB/F}$, and $N_{AB/F}$ values are 40% better than Octree, indicating the IFCNN causes less artificial noise during the combining, and it has a better information transfer rate than Octree. Thus, results show that IFCNN overcomes the limitations of the Octree method, and IFCNN's seismic attribute combination quality exceeds that of the Octree method.

V. CONCLUSION

In this paper, a method based on Deep Learning, designed to solve the problem of combing multiple seismic attributes by using CNN models is presented. The proposed method uses pre-trained general image fusion models to fuse and combine multiple seismic attributes. The approach has shown that it can:

1) Overcome the issue of lack of labeled datasets where their ground truths are identified.

2) Reduce the time and resources requirements by the use of pre-trained models to eliminate the training phase.

3) Provide a refined solution that is capable of producing better results than using the DeepFuse model.

The experiments that were conducted led to a number of findings. They showed that the Image-fusion Framework using the CNN (IFCNN) model was the best model out of all of the investigated models for combining seismic attributes. The Zero-Learning model is a lightweight model, and the easiest one to modify, and can be extended for future research to further study these types of fusion models to use for seismic attribute combining tasks. General image fusion models exhibited excellent results and showed great potential. The results showed that IFCNN scored better than other DL models on multiple metrics. Its $Q_{AB/F}$ and *MS-SSIM* scores are 50% and 10%, respectively, higher than the second-best model. When IFCNN is compared to the current state-of-the-art, Octree, IFCNN combining quality was superior to Octree especially when the number of attributes is high. Metrics results showed that IFCNN is better the Octree by 40% when the two are compared by the quality of the image structural information and the amount of noise.

The work presented in this paper implemented a method that can combine seismic attributes using pre-trained DL models. The choice of the pretrained models was based on their performance in comparison to other models found during the literature review. With scientific research advancement, better models can be introduced. The downside to the study presented is that the work is limited to seismic attributes used for faults enhancement and detection.

ACKNOWLEDGMENT

We would like to extend special thanks of gratitude to Dr. Saleh Aldossary for this golden opportunity to work on this project. We acknowledge and appreciate his unrelenting effort in advising us and providing test data for the project.

REFERENCES

[1] Al-Shuhail, S. A. Al-Dossary, and W. A. Mousa, Seismic Data Interpretation using Digital Image Processing. 2017.

[2] E. L. Galvan Aguilar, "Log property mapping guided with seismic attributes," Sep. 2013.

[3] S. S. Manral and D. Clark, "Multi Attribute Analysis-An effective visualization & interpretation technique," 2010.

[4] S. Al-Dossary, J. Wang, and Y. E. Wang, "Combining multiseismic attributes with an extended octree quantization method," Interpretation, 2019, doi: 10.1190/int-2018-0099.1.

[5] B. Meher, S. Agrawal, R. Panda, and A. Abraham, "A survey on region based image fusion methods," Inf. Fusion, vol. 48, pp. 119–132, 2019, doi: 10.1016/j.inffus.2018.07.010.

[6] B. Ashalatha and M. B. Reddy, "Image fusion at pixel and feature levels based on pyramid imaging," in 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2017 - Proceedings, 2017, pp. 258–263, doi: 10.1109/ICSTM.2017.8089164.

[7] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," Pattern Recognit. Lett., vol. 23, no. 8, pp. 985–997, 2002, doi: https://doi.org/10.1016/S0167-8655(02)00029-6.

[8] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," Inf. Fusion, vol. 33, May 2016, doi: 10.1016/j.inffus.2016.05.004.

[9] A. Alotaibi, M. Fadel, A. Jamal, and G. Aldabbagh, "Enhancement of 3D Seismic Images using Image Fusion Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 1, pp. 365–372, 2021, doi: 10.14569/IJACSA.2021.0120143.

[10] E. Blasch, Y. Zheng, and Z. Liu, Multispectral Image Fusion and Colorization. 2018.

[11] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," Inf. Fusion, vol. 42, no. September 2017, pp. 158–173, 2018, doi: 10.1016/j.inffus.2017.10.007.

[12] F. Lahoud and S. Süsstrunk, "Fast and Efficient Zero-Learning Image Fusion," pp. 1–13, 2019, [Online]. Available: http://arxiv.org/abs/1905.03590.

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv 1409.1556, Sep. 2014.

[14] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," Inf. Fusion, vol. 54, no. July 2019, pp. 99–118, 2020, doi: 10.1016/j.inffus.2019.07.011.

[15] K. R. Prabhakar, V. S. Srikar, and R. V Babu, "DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4724–4732, doi: 10.1109/ICCV.2017.505.

[16] H. Yan, X. Yu, Y. Zhang, S. Zhang, X. Zhao, and L. Zhang, "Single Image Depth Estimation With Normal Guided Scale Invariant Deep Convolutional Fields," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 1, pp. 80–92, 2019, doi: 10.1109/TCSVT.2017.2772892.

[17] P. Jagalingam and A. V. Hegde, "A Review of Quality Metrics for Fused Image," Aquat. Procedia, vol. 4, pp. 133–142, 2015, doi: https://doi.org/10.1016/j.aqpro.2015.02.019.

[18] K. Vasudevan, D. Eaton, and F. A. Cook, "Adaptation of seismic skeletonization for other geoscience applications," Geophys. J. Int., vol. 162, no. 3, pp. 975–993, Sep. 2005, doi: 10.1111/j.1365-246X.2005.02704.x.

[19] Y. Luo, W. G. Higgs, and W. S. Kowalik, "Edge detection and stratigraphic analysis using 3D seismic data," in SEG Technical Program Expanded Abstracts 1996, Society of Exploration Geophysicists, 1996, pp. 324–327.

[20] J. L. Gómez and D. R. Velis, "Structure-oriented edge-preserving smoothing in the frequency domain: Application to enhance 3D seismic data volumes," in 2017 XVII Workshop on Information Processing and Control (RPIC), 2017, pp. 1–5, doi: 10.23919/RPIC.2017.8211614.

AUTHORS' PROFILE

Abrar Alotaibi is a postgraduate student in the Computer Science Department, Faculty of Computing and Information Technology at King Abdulaziz University. She received a bachelor's degree from Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, in 2015. She holds a Teaching Assistant position at the College of Computer Science and Information Technology in the Computer Science Department at Imam Abdulrahman Bin Faisal University. Her research interests are Image Processing and Machine Learning. She is a member of IEEE.

Dr. Mai A Fadel is a lecturer in Software Engineering at the Department of Computer Science, King Abdulaziz University. Her research interest lies in Design Patterns, and software engineering in general, cloud computing and news credibility in social networks and High-Performance Computing (HPC). She has taught courses in software engineering, web development, algorithms and Java programming. Dr. Mai was the Head of the Information System and the Computer Science departments between 2009 and 2016. Dr. Mai received her PhD degree in Computer Science from the Department of Computer Science at the University of Exeter, UK. She is a member of IEEE and ACM.

Dr. Amani Tariq Jamal is an assistant professor at the Computer Science Department, Faculty of Computing and Information Technology at King Abdulaziz University. She received her Master and Ph.D. degrees from Concordia University, Montreal, Canada, in 2009 and 2015 respectively. Her research interests are pattern recognition, artificial intelligence, Natural Language Processing, and document analysis. She is the cofounder and member of the directors' board of Saudi AI Society.

Dr. Ghadah Aldabbagh received her BSc in 1991 in Computer Science, from the College of Engineering at the University of Illinois at Urbana – Champaign, Illinois, United States; she received her MSc in the Data Communication Networks and Distributed System (DCNDS) from the Computer Science Department at the University College London (UCL) in the United Kingdom; received her PhD in 2010 from the Department of Electronic and Electrical Engineering at UCL, UK. Dr. Aldabbagh holds an Associate Professor position at the Computer Science Department at the Faculty of Computing and Information Technology (FCIT) at King Abdul Aziz University (KAU), Jeddah, Saudi Arabia. Her current research interests: 5G, Wireless Communications, Wireless Sensor Networks, Internet of Things (IoT) Platform, Internet of Everything (IoE), Connected Homes, Smart Workspace, Machine Learning, Deep Learning, Deep Reinforced Learning, Artificial General Intelligence, Artificial Intelligence Everywhere, Autonomous Vehicles, providing connectivity of IoT devices and sensors for Smart Cities, Location Awareness. During the summers of 2013 and 2014, Dr. Aldabbagh joined Professor John M. Cioffi's Dynamic Spectrum Management (DSM) Research Group as a visiting professor at Stanford University. She was awarded by KAU four awards of excellence in scientific publishing for faculty members for the academic year 2015-2016. In March 2016, Dr. Aldabbagh joined the Laboratory for Information and Decision Systems (LIDS) at Massachusetts Institute of Technology (MIT) as a postdoctoral associate and from March 2017 as a visiting scholar to work on the topic of location-aware ad-hoc sensor networks to work with the highly cited Professor Moe Win. In collaboration with Professor Win's research group, her focus is to develop scalable, distributed, and energy-efficient techniques for scheduling and routing in ad hoc sensor networks with localization-awareness. In November 2019, she participated in the Saudi Think Tank 20 (T20) of the Saudi G20, as a co-author for two policy briefs (PB); the first was in Task Force 10 (TF10): Sustainable Energy, Food and Water Systems and the second PB was TF6: Economy, Employment, And Education In The Digital Age. In April 2020, Dr. Aldabbagh joined the Marconi Society as a member of the Marconi Selection Advisory Committee for the Marconi Prize.

Computational Intelligence Algorithm Implemented in Indoor Environments based on Machine Learning for Lighting Control System

Mohammad Ehsanul Alim[1]
Department of Electrical & Computer Engineering
University of Delaware, Newark, Delaware, USA

Md. Nazmus Sakib Bin Alam[2]
Department of Electrical & Computer Engineering
North South University, Dhaka, Bangladesh

Sneha Shrikumar[3]
School of Electrical & Electronic Engineering
Nanyang Technological University, Singapore, Singapore

Ihab Hassoun[4]
Faculty of Engineering
City University, Tripoli, Lebanon

***Abstract*—Over the past decade, engineers and scientists dedicated a significant amount of effort and time to enhance an indoor system embedded with the state-of-the-art automation. Through innovative implementation of sensors, IoT and machine learning algorithm, the designing of indoor lighting control systems evolved over the period. Our research is based upon the development of a highly intelligent lighting system that will be cost effective and at the same time easily accessed in a remote mode. Devices like Ultra-wide band sensors and Lux sensors were collected and utilized in the designing of the system to retrieve information about the user's location and existing brightness in the room, respectively. These data were then preprocessed, scaled and transmitted to various machine learning algorithms to predict suitable lighting condition. The application of our proposed lighting system will always keep the brightness range to a recommended level of 200-400 Lux which is extremely compatible for its use in homes, offices, schools and high rage apartments. In addition, the remote access facility allows users to operate the system anywhere in the world providing user experience beyond imagination. Lastly, as the system comprises of low-cost components that are also easily replaceable and only provide lighting when needed, it can provide savings in terms of cost and power.**

***Keywords*—Machine learning algorithms; indoor lighting control system; internet of things (IoT); ultra-wide band sensors; lux sensors; remote access facility**

I. Introduction

The importance of designing an efficient indoor lighting system is impeccable in modern times. Interestingly, the priority of lighting system is not only circumscribed to exterior work but also light is a pivotal entity in terms of its significance to human health. From the production of Vitamin D- an essential ingredient for the bone development to Vitamin D2 for conversion of ergosterol, light contributes to such important functions in human health. Nowadays light even being used as a therapy for numerous diseases ranging from sleeping disorder to Alzheimer due to the fact that light always triggers visual cortex of the brain used for vision and other neural activities.

Undoubtedly, the admittance for the advantages and significance of productive lighting is inevitable [4]. This is why continuously engineers and scientists all around the world try to develop intelligent indoor lighting system with the application of machine learning. We know that machine learning deals with developing algorithms that can gain access to large amount of data and learn automatically both in the form of supervised and unsupervised mode. In addition, due to the extensive application of Internet of Things (IoT) like wireless technology and sensors, it is extremely viable to connect devices with the internet and exchange large amount of data. For example, it has been reported that the combination of IoT along with machine learning technology will lead to an increase of worldwide market earnings from $651 million in 2017 to $4.5 billion in 2026 [24].

The implementation of sensors with lighting makes the overall use of control hardware redundant. Luminaries can communicate by exchanging data with one another, allowing seamless change in lighting to suit the immediate environment with manual intervention. Interestingly, the intervention of mobile apps without installing any additional cables makes the indoor lighting system cost effective, secured, sustainable, user friendly and most importantly desirable to consumers and flexible for other technologies as well [11]. Generally, when IoT solutions are incorporated, high expenditure for suitable infrastructure is incurred, making return on investment uncertain. However, with intelligent lighting, lesser energy consumption is certain, reducing power costs. By further tuning and daylight harvesting, power savings by up to 90% are possible through LED lighting [12].

In this section, the need for good indoor lighting has been discussed. In the next section, the existing research on intelligent lighting has been reviewed. Further, the latest research on the impact of incorporating intelligent lighting has been highlighted in proposed methodology section. Finally, the various intelligent lighting systems that have been built lately have been examined in the last section.

II. Literature Review

The importance of intelligent lighting system is something that is taken into serious consideration by many and as a consequence we have seen some prolific work that already proved to be productive.

Light is an essential aspect for both functional and physiological facets of human beings. Some of the examples include its importance in terms of the production of melatonin which is a hormone that prepares the body for sleep at night [1]. Importance of sleep to human physiology is beyond articulation as several vital functions of brain – cognition, concentration and productivity is directly related to sleep [2]. Needless to mention the significance of light from an employee's perspective as inadvertent designing of lighting can result mishaps, lethargy, annoyances and most unwanted accidents. ILO Manual states that improved lighting can increase productivity up to 10% and reduce error by 30%.

When the lighting in an area adapts to changes in its surroundings to achieve the necessary illumination, intensity, and user satisfaction, it is called intelligent. It is proven fact that intelligent lighting can result a sustainable and effective increase in productivity and performance. Research has shown that available lighting in the office directly affects employee's productivity, performance and probability of causing accidents [3]. Considering the overall market segment and growing focus on achieving a healthier and comfortable life, intelligent lighting is one of the fastest growing areas for research and commercial and probably it will not be wrong to term it as the future of lighting. The process of installing intelligent lighting in every household is much easier now due to an enormous number of devices around 22.5 to be precise connected to Internet by the end of the year 2021 based upon the Ericsson's latest Mobility Report.

While lighting control has been commercially prevalent over the last few decades, the ways to achieve lighting control are constantly evolving. At the very beginning, lights could only be controlled using switches and dimmers. Users had to manually change the light settings to suit their preferences. As sensors and microcontrollers were commercially made available, they could replace traditional switches by changing brightness and intensity levels without user intervention [5]. Furthermore, by incorporating memory devices in lighting systems, personalized lighting control has become possible by saving user's preferences [6]. In today's times, easy availability of data and data mining technologies has made intelligent lighting possible. In intelligent lighting, machine learning algorithms are used to learn from available data or through their interaction with the environment to achieve optimal light settings without manual intervention. The evolution of lighting control over years has been summarized in the below Table I.

TABLE I. Advancement of Lighting Control Algorithms

Type of Lighting	Extent of User Intervention	Behavior	Commercial Example
Traditional	Greatest	One has complete control over lights	On or Off switches
Autonomous	Null	Lights need not be controlled by users	Controlled by sensors
Adaptive	Least	Stores user's preferences	Lighting is personalized using memory devices
Intelligent	Minimum	Lighting systems learns user preferences	Intelligence lighting control systems using machine learning

There are several intelligent lighting algorithms that have been proposed in recent times to achieve certain objectives by changing their behavior. Some of these systems include:

- Indoor Light Automatic Control System (ILACS) control Algorithm.
- Structure of lighting control system using evolutionary optimization algorithm.
- Intelligent dimming using PIR sensor and a dimmer circuit.
- Intelligent room using Fuzzy Logic.
- Using power line carrier design.

A. Indoor Light Automatic Control System (ILACS)

Prominent researchers [8] have proposed an Indoor Lighting Automatic Control Algorithm (ILACS) that controls lighting according to daylight intensity, occupancy and motion detection to ensure efficient utilization of energy. This work focuses on Radial Basis Function Neural Network & Generic Algorithm [10].

When operating on ILACS, the algorithm works as per the selection of either of the two control methods- Lighting control by scheduling and lighting control as per occupancy detection. The two methods have been illustrated in the Fig. 1.

When the former control method is selected, the algorithm calculates the daylight illuminance as per the present time zone. The required light level is then determined based on the light sensor and daylight calculations. The algorithm then regulates the lighting groups in the system using light patterns in accordance with the determined light level.

This system has been implemented using a light and PIR sensor to detect illuminance and occupancy and a Linux system to run the ILACS algorithm. The lighting control information is transmitted wirelessly through ZigBee modules to the various Lighting groups in the system- consisting of LED drivers, down light and flat light by following closed loop control algorithm [7].

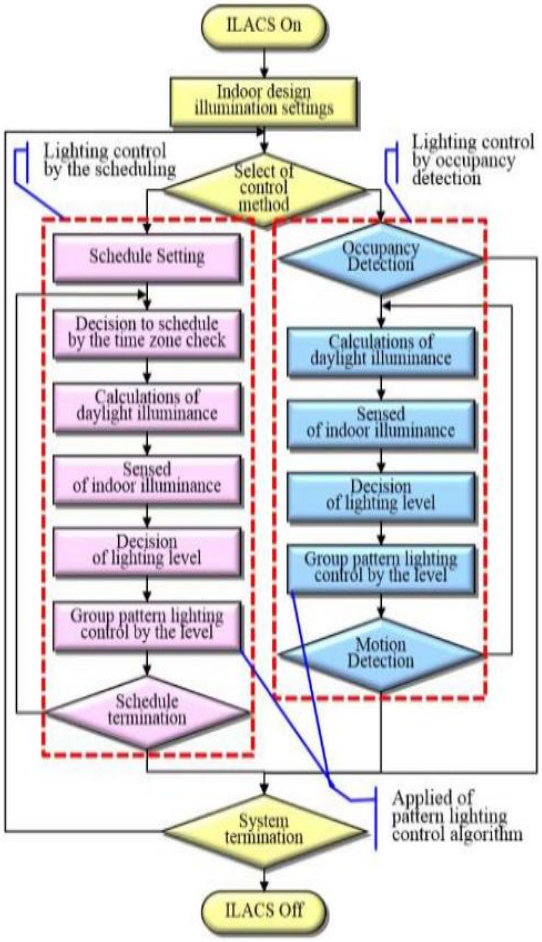


Fig. 1. ILACS Control Flow.

Results of this set up state that using the ILACS algorithm utilizing external daylight with LED lighting reduces the power consumption by up to 66.25%.

B. Evolutionary Optimization Algorithm

The researchers in this work elaborated the design and implementation of a lighting control system that is built on genetic algorithm optimization (Fig. 2). It emphasizes on maximum utilization of natural daylight to achieve user comfort and energy efficiency. The luminaire is either dimmed or switched on/off as per the measured horizontal and vertical illuminance values in the room. Using the genetic algorithm, the illuminance output of all the lighting devices in the workspace is estimated and the optimum operating schedule of the electric system is decided.

The proposed system is an easily implementable and inexpensive solution that offers efficient energy consumption by reaping the benefits of natural daylight. Lights can be switched on or off manually, but not dimmed. In the work by I. Petrinska, V. Georgiev and D. Ivanov [9], the design and implementation of a lighting control system that is built on genetic algorithm optimization has been detailed which is same like our proposed system.

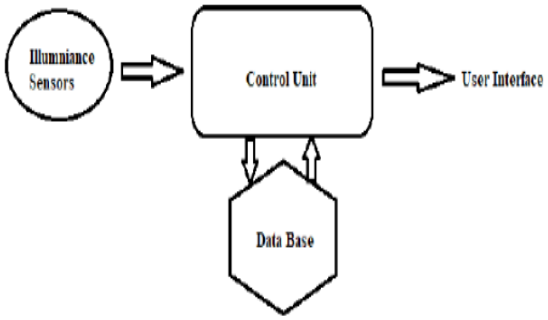


Fig. 2. Evolutionary Optimization Algorithm Implemented in Lighting.

C. Intelligent Dimming using PIR Sensor and Dimmer Circuit

The researchers implemented a dimmer circuit in the design that varies the Pulse Width Modulation (PWM) signals according to user's position. The whole system can be remotely operated using LabVIEW software.

When the PIR sensor detects motion, it sends a high value signal to the Arduino Uno microcontroller. The microcontroller then generates the PWM signal that is transmitted to the dimmer circuit to switch on the light with maximum brightness. Similarly, if no use has been found, the PIR sensor transmits a low voltage signal to switch the LED off.

The proposed system (Fig. 3) has applications in many other fields like street lightings, industries, and homes. While it is a low cost set up that can potentially save power, it does not utilize smart algorithm to control output voltage and hence it is not fully intelligent.

D. Intelligent Room using Fuzzy Logic

The researchers in the designing of intelligent room using Fuzzy Logic System have simulated a smart lighting room that takes the room temperature and available lighting into consideration to produce necessary voltage [13]. This voltage is then used for LED and blind control by using fuzzy logic.

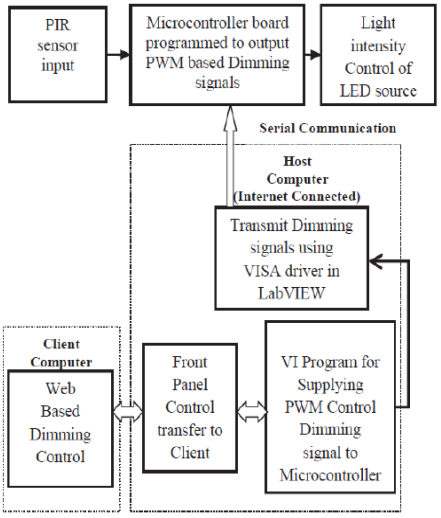


Fig. 3. System Block Diagram for Intelligent Dimmer Circuit.

The set up utilizes occupancy and lux sensors to detect presence of user and the present illumination levels, respectively. The sensitivity of the sensors must be such that it neither produces delay in delivering information nor consumes excess power. An additional sensor called radar motion sensor is also used to alert the system to consume less power when there is no activity in the room beyond a certain duration. Window blinds with actuators are used in this experiment to automatically allow maximum outdoor light into the room while preventing glare.

However, system using fuzzy logic requires complicated and enhanced computation power. Therefore, replacement of fuzzy logic by machine learning algorithms has been suggested.

E. Power Line Carrier Design

The researchers in this proposed system [14] have introduced an intelligent control using power line communication and a GSM module for remote control. The system design is illustrated using Fig. 4.

The microcontroller communicates with the GSM module via an RS232 cable for serial communication. It is also serially connected to the power line for communication. Effective transmission between the nodes is carried out by the power line cattier module. For encoding and decoding purposes, PT2248 chip is used. The infrared circuit in the chip receives, detects, modulates and demodulates the infrared signal. Lighting brightness control is achieved by the lighting terminal controller.

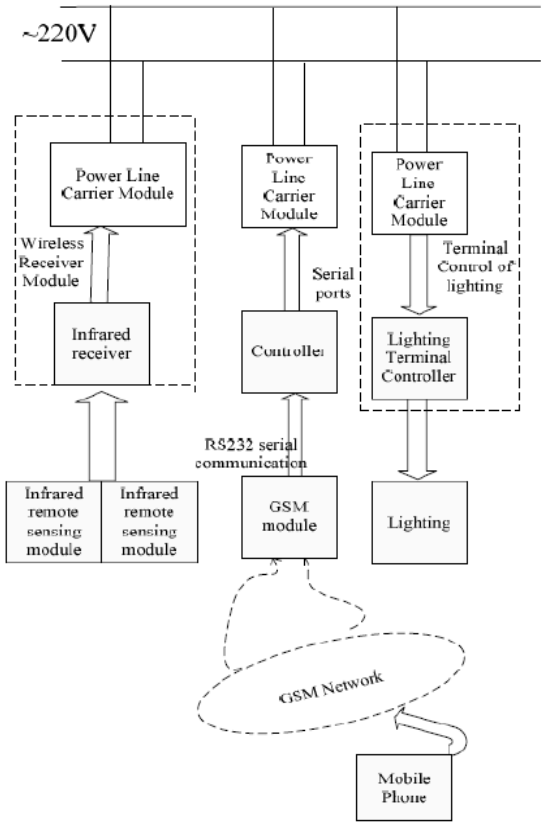


Fig. 4. System Design.

III. Proposed Methodology

A. Proposed Experimental Procedure

The primary criteria of the implementation is to provide suitable lighting conditions in accordance with the immediate environment. The various aspects that form the immediate environment will be the features impacting the project. By understanding these features better, the data that must be collected and fed into the algorithms for providing a suitable output can be determined. These features depend on the architecture of the area in which the project must be set up.

Therefore, the layout of the room where the project has been set up comes into consideration. In order to understand the nature of intelligent lighting, information about the various factors influencing lighting in the area and the relationship among these factors are crucial. This information will help understand the problem statement better that will enable in the design of a suitable lighting system. Since the primary criteria of the implementation is to provide suitable lighting conditions in accordance with the immediate environment, the various aspects that form the immediate environment will be the elements (features) impacting the project. By understanding these features better, the data that must be collected and fed into the algorithms for providing a suitable output can be determined. These features depend on the architecture of the area in which the project must be set up. Therefore, the layout of the room where the project has been set up comes into consideration. The layout of the room where the experiment was set up has been shown in the below Fig. 5.

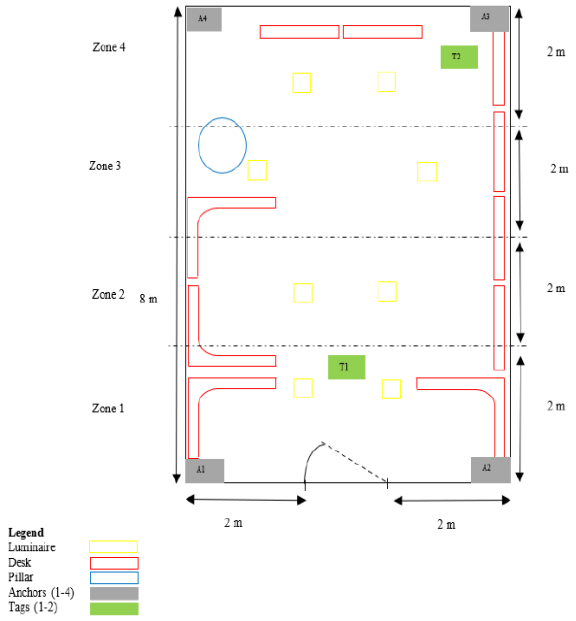


Fig. 5. Experiment Set up.

Experiment set up consists of 8 LED luminaires that are controlled by Mi-Light 40W 0/1 ~ 10V Drivers to switch the lights on and off or for dimming purpose [15]. To ensure even distribution of light throughout the room, the room has been divided into 4 zones accordingly as shown in the above figure. Therefore, the dimension of each zone is 5.5m x 2m x 2.2m. Each of the zones is operated by a set of two luminaires,

controlled by the Mi-Light drivers. Mi-Light 40W 0/1 ~ 10V Drivers are used to adjust the brightness of the lights from 0 to 100% via PWM signals. The controller uses 2.4 GHz wireless technology for low power and wireless transmission. Wireless connection to the driver can be established via the Mi-Light app in a smartphone or through the Mi-Light iBox2 Wi-Fi bridge [17]. By linking this bridge to the local Wi-Fi router, remote access to these lights becomes possible from anywhere in the world. To detect real time positions of users, 4 DecaWave DWM1001 RTLS units are configured as anchors and mounted at every corner of the room. Each of these anchors are placed 1.5 m above floor level to guarantee line of sight for better performance and seamless operation. Some of the specifications of Mi-light driver [16] are given in Table II.

Two of the remaining RTLS units are then configured as tags to record and mimic users' position in the room. Furthermore, to provide suitable lighting conditions to users in each zone, information about the existing lux level in each of the zones is necessary. For this purpose, EDS OW-ENV-THPL sensors are mounted at the center of each of the zones. These sensors also capable of providing temperature, humidity and barometric pressure reading besides lux levels. However, to ensure relevancy to the experiment, only lux levels are recorded at the rate of 1 reading per second. This data can be acquired via a data aggregator called MeshNet controller from which data can be stored and viewed at real time. This useful data from both the anchor tags and lux sensors is then fed to Raspberry Pi 3 for manipulation by the algorithms. The algorithm then accordingly predicts a suitable lighting condition as output based on the data given which is then fed back to the Mi-Light bridge to adjust the lights in the zone accordingly.

The MDEK1001 kit is used to create an RTLS (Real time location system) that provides data about the location of the user in the room [18]. The RTLS was created by initializing 4 DWM1001 development boards in the kit as anchor nodes, 1 board as a bridge and 2 boards as tags via the Decawave DRTLS Manager android application [20]. The anchors were mounted on the 4 ends of the wall to create a UWB network and compute the location of the mobile 'tags' in the room. To transmit information from the UWB network to Raspberry Pi's IP network (for instance LAN) and vice versa, a separate board was configured as a bridge. It collects location information from the anchors in the network and transfers them to the Linux operating system of Raspberry Pi in the network. Therefore, by using the bridge, monitoring the UWB network via an external network was possible [19]. This configuration is possible either by using the smartphone DecaWave RTLS Manager application via Bluetooth, or by a desktop or microprocessor via an SPI or UART connection. The Tags were powered by rechargeable batteries, the bridge by the SPI connection to the microprocessor and the anchors by USB power supplies [20]. The specifications of the development boards are summarized as follows [20] in the below Table III.

The MDEK1001 kit is used in this project to create an RTLS (Real time location system) that provides data about the location of the user in the room. The RTLS was created by initializing 4 DWM1001 development boards in the kit as anchor nodes, 1 board as a bridge and 2 boards as tags via the Decawave DRTLS Manager android application [20]. The anchors were mounted on the 4 ends of the wall to create a UWB network and compute the location of the mobile 'tags' in the room. To transmit information from the UWB network to Raspberry Pi's IP network (for example- LAN) and vice versa, a separate board was configured as a bridge. It collects location information from the anchors in the network and transfers them to the Linux operating system of Raspberry Pi in the network. Therefore, by using the bridge, monitoring the UWB network via an external network was possible [19]. This configuration is possible either by using the smartphone DecaWave RTLS Manager application via Bluetooth, or by a desktop or microprocessor via an SPI or UART connection. The Tags were powered by rechargeable batteries, the bridge by the SPI connection to the microprocessor and the anchors by USB power supplies [20]. The specifications of the development boards are summarized as follows [20] in the above Table III.

On successful configuration of the RTLS, the position of the tags in the room with length and breadth as 8 m and 5.45 m bounded by the 4 anchors were visible in the app as shown in the screenshot (Fig. 6).

TABLE II. MI-LIGHT PL1 40W 0/1~10V DIMMING DRIVER SPECIFICATIONS

Parameter	Specification
Input Voltage	AC 180-240V (50/60 Hz)
Output Voltage	DC 30-40V
Output Power (Max)	40 W
Dimming Range	0-100 %
Remote Control Distance	30m
Output Current (Constant Current)	900mA

TABLE III. TABLE SPECIFICATIONS OF DWM1001 MDEK1001 KIT [20]

Parameter	Specification
X-Y Location Accuracy	<10 cm by LOS
Normal Update Rate	100 ms / 10 Hz
Stationary Update Rate	100 ms / 10 Hz
Flash Memory available to user	40 kB
RAM Memory available to user	5 kB
Data throughput from tags to bridge	340 bytes per second (Uplink or Downlink)
Data throughput anchors to bridge	34 bytes per 12 sec (Uplink or Downlink)
System Latency	100 ms
UWB Channel	6.5 GHz (Channel 5)

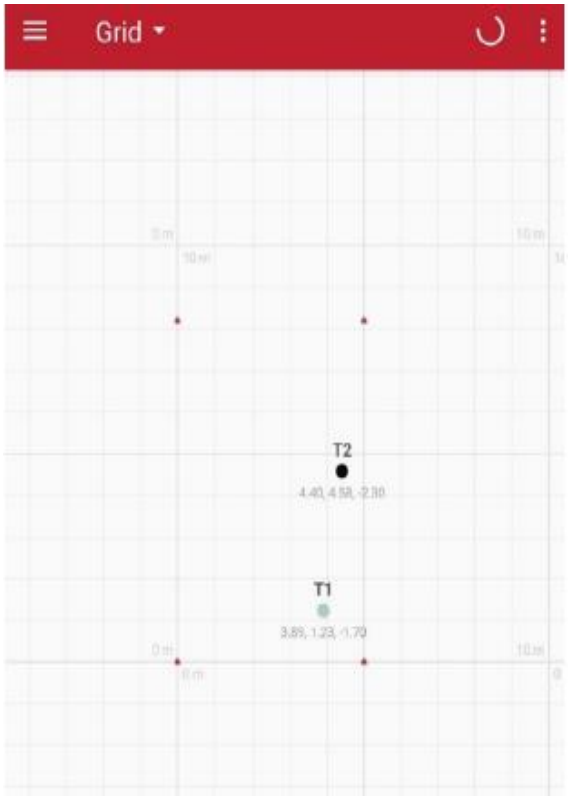


Fig. 6. DRTLS Application showing the Location of the 2 Tags in the Room.

In order to provide fitting lighting conditions in each of the zones, data about the existing lux level in each zone is necessary. For this purpose, 4 OW-ENV-THPL sensors were used in the set up to monitor the brightness levels. Each zone has 1 sensor that was mounted in the center of the desk to gather information about the available lux at desk level. The sensor has an inbuilt adapter that uses 1-Wire communication protocol to read the sensor data and convert it into readable information [21]. To transmit this information to the outside world, a MeshNet controller is used. It acts like a gateway that connects the sensor network with the external network. The controller also provides a web interface which can be used to view real-time sensor data for administrative purposes and also for sensor network configuration.

The Raspberry Pi 3 is an integral part of our research paper since it is used for a variety of reasons: for collecting data from sensors, programming and running the algorithms and transmitting commands representing the output lighting condition to the Mi-lights. The microprocessor was chosen for implementation because it offers High processing power, Ease of portability, Raspbian OS supports Linux commands that are crucial for exchanging data with the DecaWave sensor, USB compatible, Wireless LAN and Bluetooth connectivity. Some important specifications of Raspberry Pi Model 3 are given in Table IV [22].

To develop a functional lighting control system, all components must be successfully integrated via a computer program in Raspberry Pi that enables exchange of data with one another as Fig. 7. The program must carry out various actions in the sequence which are- Firstly, read data from the UWB sensor bridge iteratively to detect if user is present in any of zones in the room. Secondly, if occupancy is detected in any of the 4 zones, retrieve lux information of the occupied zone(s) from the lux sensor(s). Thirdly, feed this lux data as input to the trained algorithm to predict output brightness of the zone(s). Then, send this predicted value as command to the Mi-light bridge to change the brightness accordingly. Finally, store the number of occupied zones and the corresponding predicted output brightness is in the database.

TABLE IV. RASPBERRY PI MODEL 3 SPECIFICATIONS

Parameter	Specification
CPU	Quad Core 1.2 GHz Broadcom BCM 2837 64bit CPU
RAM	1 GB
Protocol	100 Base Ethernet, 2.4 GHz 802.11n wireless
USB Ports	2
Power Input	5V/ 2.5 A DC
Bluetooth	4.2 BLE

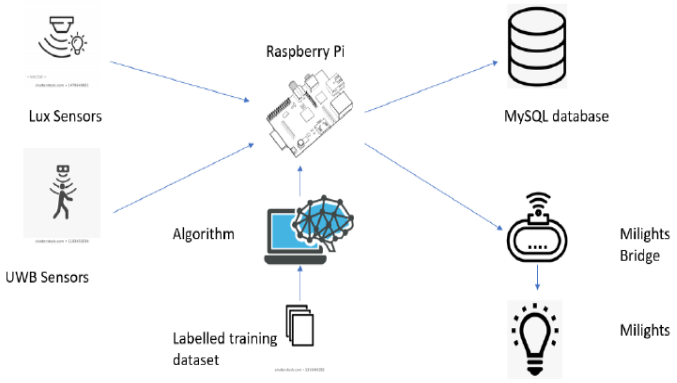


Fig. 7. Lighting Control System.

B. Experiment Execution Procedure

The selection of an algorithm for an application is based upon the understanding of its nature, features of application and expected outcome. According to the set up shown in Fig. 8, the sensor data was retrieved via Python 7 program in Raspberry Pie. The n it was stored in MySQL database during office working hours for a week's duration. Amount of light that was available at desk level was recorded by the Lux sensors at every 4 seconds. The available light was variable in nature as the zones were subjected to different number of lights at different timings of the day. The stored MySQL data contained 9455 samples that were imported to a csv file and then analyzed. The lux values varied from time to time, with the least value being 0 lux when no light is available and most being 733 during maximum light.

As per the indoor lighting recommendations in country like Singapore, the minimum light level to ensure health and safety for users at any area at the office is 100 lux and average lighting at desk level is 200 lux [23]. There are severe side

effects on exposure to high lux levels like headaches and eye strain. Our research project tried to keep the level of lux keeping in mind the visual comfort level which ranges from 180-400 lux. Hence to avoid fatigue and comply with the health and safety requirements set by the lighting standards of Singapore, the proposed algorithm must provide lighting conditions that ensure lux levels in the target range. The output of the algorithm must be either of the 5 brightness commands- 0%, 25%, 50%, 75% or 100% to the Mi- Light drivers. The flow of the Python program including the outputs of the algorithm to be implemented were accordingly developed as shown in Fig. 8.

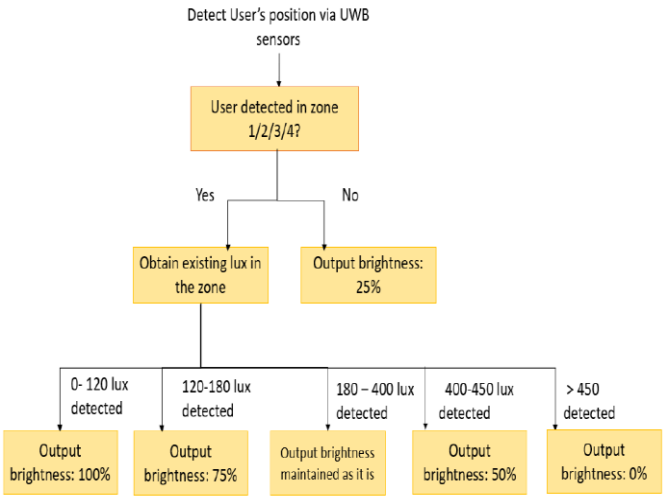


Fig. 8. Program Flow.

The program reads data from UWB sensors iteratively. If occupancy detected in any of the 4 zones, it collects the lux data from the occupied zones. If the lux level is less than or equal to 120 lux, the brightness of the light enhanced by 100%. The system boosts the level of light by 75% when the lux level is within 120 to 180. When the lux range is within the range of 180-400, the brightness level will be maintained as it is. The brightness of the light is reduced by 50% when the lux level reaches between 400 to 450 lux. However, if the lux level is more than 450, this means the existing lighting conditions are sufficient and indoor lightning is not required in this zone. Therefore, the lights are switched off by the system. The default brightness vale of 25% will be maintained for zones where there is no occupancy. The reliability of accurate data is immense and in order to eradicate flawed data due to external disturbance like noise, the importance of data preprocessing is imperative. Our proposed system implements the process of data preprocessing using Python 3.7. The collected data first imported in libraries. The programming in Python 3.7 allowed us to import datasets, look out for missing values, replacing categorical values, splitting the datasets inti training sets and performing scaling features.

Our proposed system of indoor lighting system used machine learning algorithms to establish a correlation between the situation and the target lighting settings. Basically, there are three approaches of machine learning-supervised, unsupervised and reinforcements learning algorithms. Supervised learning algorithms map the key features with the target variables using training data containing examples that include vectors and their corresponding outputs [25]. For training data containing input variables without outputs, unsupervised learning algorithms are beneficial to find the exact relationship in the data [26]. The model uses feedback from which it can learn and is comparable to supervised learning. However, this feedback is usually noisy and includes delay that can make it difficult for the model to establish a connection with the inputs and outputs [27]. Once trained, the algorithm will establish a hypothesis that predicts the output lighting settings based on the new inputs [8]. In our proposed system, the output must be either of the 5 output light settings: 0%, 25%, 50%, 75% or 100% brightness. This is why for successful implementation of the system, application of classification supervised learning algorithm is necessary. Some of the widely used supervised learning algorithms selected for implementation include:

- Logistic Regression.
- K-nearest Neighbors.
- Support Vector Machine.
- Kernel SVM.
- Naïve Bayes.
- Decision Tree.
- Random Forest.

1) Logistic regression: Logistic regression is a classification technique from the field of statistics and probability that used to categorically describe the relationship between data. It is an extension of linear regression that can be applied for classification problems. The prediction for the output is transformed using a non-linear function called logistic sigmoid function [29].

2) K-Nearest neighbors: K-Nearest Neighbors is a straightforward algorithm which stores all training classes in a graph and classifies the incoming datapoint based on its similarity with its surrounding neighbors on the graph [31]. In this application, as k = 5, the datapoint is assigned a class that is in the majority vote of 5 of its neighbors.

3) Support vector machine: Support Vector Machines (SVM) are one of the most widely known and used machine learning algorithms known today. In this algorithm, all input points in the feature space are separated according to their class by a line called a hyperplane [30]. The objective is to discover the best coefficients that result in the most accurate partition of the points by the hyperplane. The hyperplane is also called a decision boundary in certain cases.

4) Kernel SVM: Kernel SVM is used in practice to choose a decision boundary for non-linearly separable data. This is done by taking nonlinearly separable data set and mapping it to a higher dimension to get a linearly separable data set [32]. The algorithm works as follows: invoke the support vector machine algorithm, build a decision boundary for the dataset, and then project all of that back into original dimensions. One

of the most popular kernel functions used in SVM is the radial basis function kernel, or RBF kernel. For feature vectors x and x' in the input space, RBF kernel is defined as [33]:

$$K(x, x^1) = exp\left(-\frac{\| x - x^1 \|^2}{2\sigma^2}\right)$$

$\| x - x^1 \|^2$is the square of Euclidean distance between x and x' and σ is a free parameter. Other famous Kernel functions include sigmoid, marten and Polynomial kernel.

5) Naïve Bayes: Naive Bayes is a straightforward but robust predictive modeling algorithm. It consists of two types of probabilities that are calculated from the training data. Two types of probabilities are given below:

a) The probability of every class.

b) The conditional probability for each class for each value.

Once computed, this algorithm can predict outputs for incoming data by applying Bayes Theorem. For data that is real-valued, a Gaussian distribution is generally assumed for easy probability estimation [34]. Naive Bayes is labeled as 'Naïve' because it assumes that all incoming variables are independent. However, in reality, this assumption is very unrealistic due this wrong assumption. Nevertheless, the algorithm still performs really well for a wide range of complex applications [28].

6) Decision tree: Decision Trees is another important model that is commonly used for predictive modelling in machine learning. This algorithm is essentially represented as a binary tree. Every node in the tree stands for an incoming input variable that can also be spit in the future iterations. The last nodes of the tree are referred to as leaf nodes. These nodes contain the data that the model finally used to predict an output. The algorithm runs through the entire tree via splits till it arrives at the leaf node to predict the output class value.

7) Random forest: Random Forest (also referred to as Bootstrap Aggregation or bagging) is another popular machine learning model that is based on ensemble learning. This algorithm is a robust statistical technique that can estimate a quantity from the input data using metrics like mean. The algorithm works as follows- collect many samples of input data and find the mean of each input value. Once done, find the average of all mean values to get the true mean value. In this way, models are built for every training data sample. When a new input data arrives, each of the models predicts an output value. Then, all these predictions are then averaged to find the true output for the input data.

IV. Result and Discussion

The collected dataset is pre-processed and then used to train the 7 supervised algorithms. The dataset contains 4996 samples that have been split in the ratio 75:25 for training and testing purpose. In order to evaluate each algorithm, a graph representing the actual versus the predicted values was plotted. To gather further insight about each algorithm, metrics such as confusion matrix (visualized with the help of a heatmap) and accuracy score were also generated. Accuracy score is a function in python that calculates the percentage of predicted values that accurately match the actual values of a particular label in the dataset. The higher the accuracy score, the better. Similarly, a confusion matrix is a similar metric that is used to measure model accuracy. It is a summarized matrix containing the number of correct and incorrect predictions that are broken down by each class. The rows and columns of the confusion matrix describe the true positives, true negatives, false positives, and false negatives of each class in an algorithm.

A result is false positive when the algorithm incorrectly predicts the presence of a certain class for an input value when it actually is not the case. Similarly, a false negative result occurs when the algorithm incorrectly rejects the presence of a class. On the contrary, correctly identified predictions are called true positives and righty rejected prediction are defined as true negatives. The results of each machine algorithm are as follows (Fig. 9 to 22):

1) Logistic regression

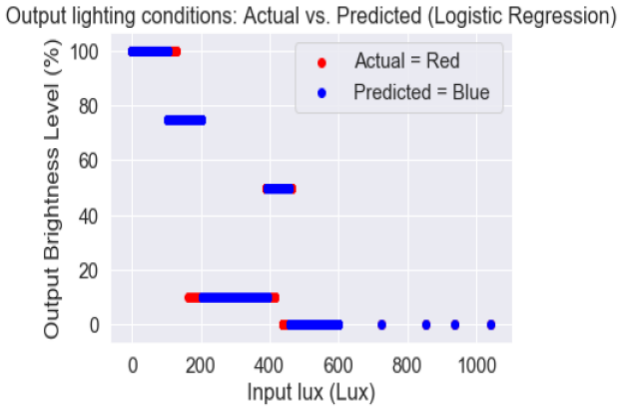


Fig. 9. Graph Comparing the Actual and the Predicted Output Values of Logistic Regression.

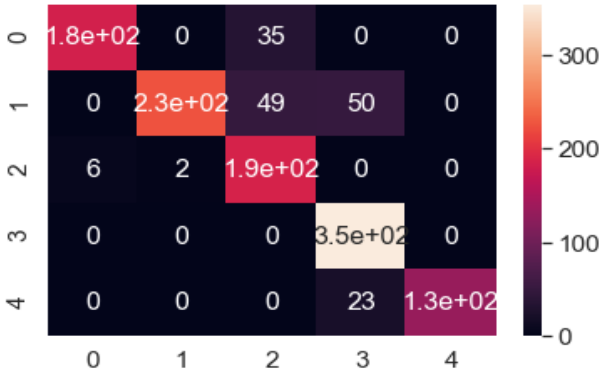


Fig. 10. Confusion Matrix of Logistic Regression.

Accuracy Score for Logistic Regression: 86.7812%.

2) K- Nearest neighbors

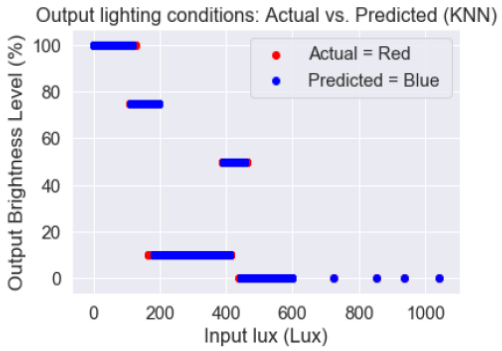


Fig. 11. Graph Comparing the Actual and the Predicted Output Values of K-Nearest Neighbors.

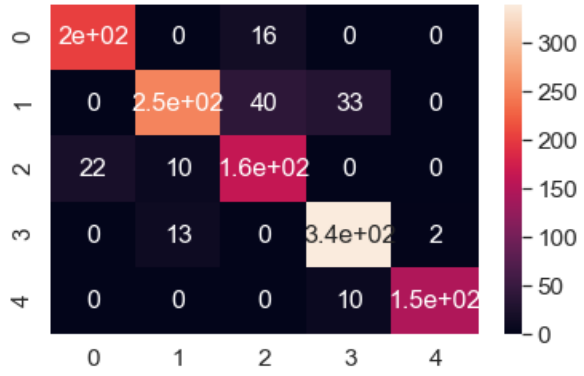


Fig. 12. Confusion Matrix of K-Nearest Neighbors.

Accuracy Score for K-Nearest Neighbors: 88.3106%.

3) Support vector machine

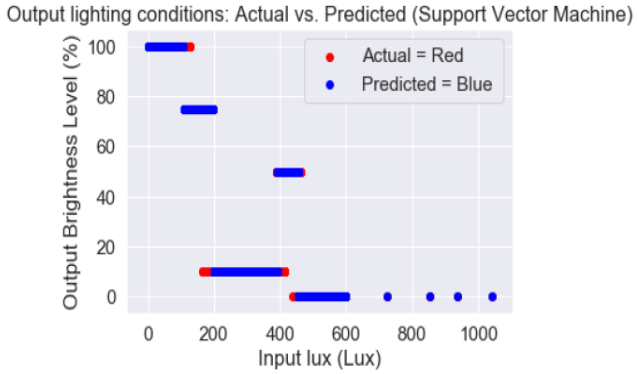


Fig. 13. Graph Comparing the Actual and the Predicted Output Values of Support Vector Machine.

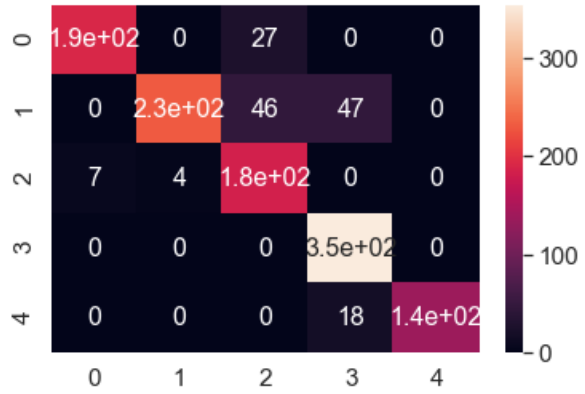


Fig. 14. Confusion Matrix of Support Vector Machine.

Accuracy Score for Support Vector Machine: 88.0704%.

4) Kernel SVM (Kernel Support Vector Machine)

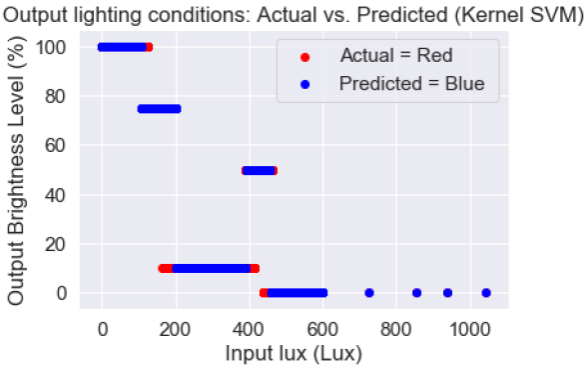


Fig. 15. Graph Comparing the Actual and the Predicted Output Values of Kernel Support Vector Machine.

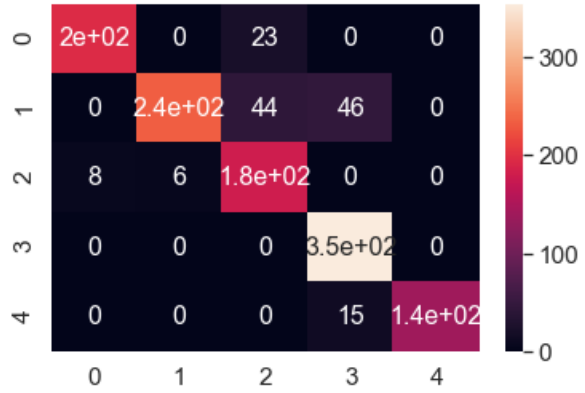


Fig. 16. Confusion Matrix of Kernel Support Vector Machine.

Accuracy Score of Kernel Support Vector Machine: 88.6309%.

5) Naïve Bayes

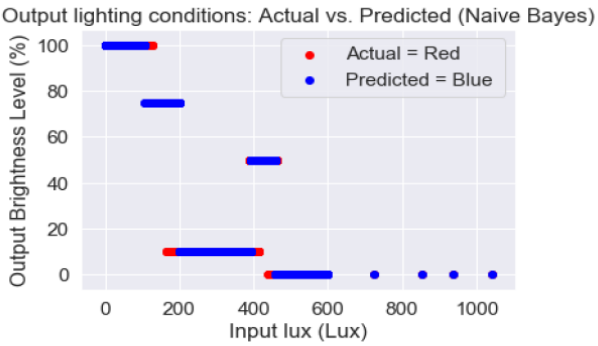


Fig. 17. Graph Comparing the Actual and the Predicted Output Values of Naïve Bayes.

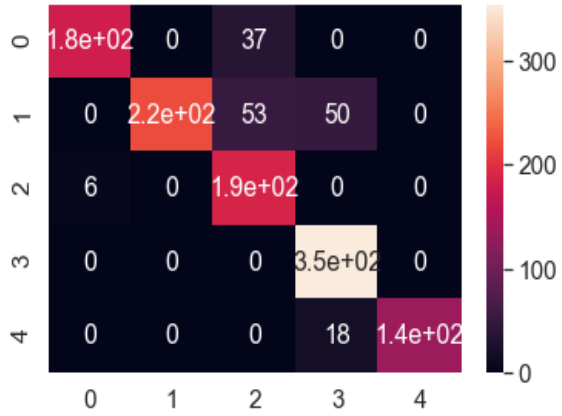


Fig. 18. Confusion Matrix of Naïve Bayes.

Accuracy Score of Naïve Bayes: 86.8694%.

6) Decision tree

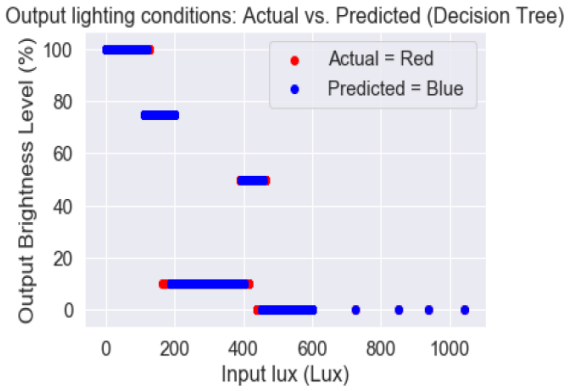


Fig. 19. Graph Comparing the Actual and the Predicted Output Values of Decision Tree.

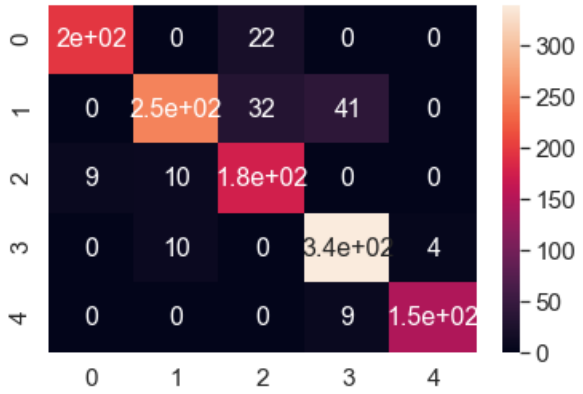


Fig. 20. Confusion Matrix of Decision Tree.

Accuracy Score of Decision Tree: 89.0312%.

7) Random forest

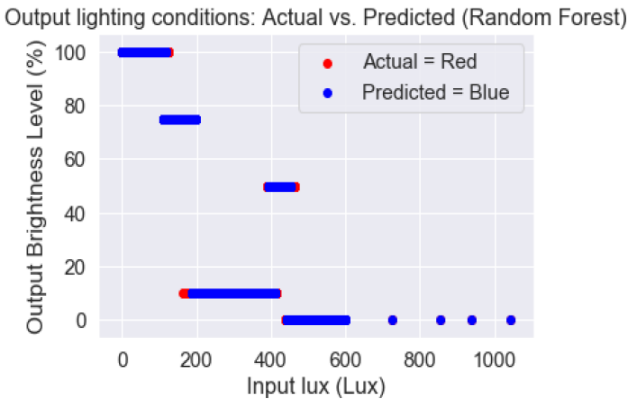


Fig. 21. Graph Comparing the Actual and the Predicted Output Values of Random Forest.

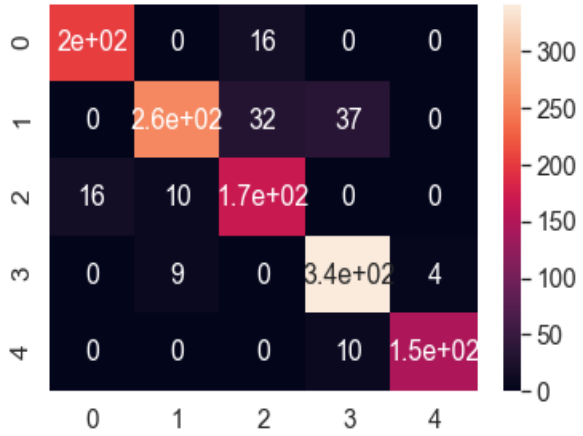


Fig. 22. Confusion Matrix of Random Forest.

Accuracy Score of Random Forest: 89.2714%.

V. Evaluation and Modification

It is evident that random forest algorithm has the highest accuracy score. This result is found from the above algorithm. If the provided dataset contains equal number of samples of each class, for instance, if it is a balanced dataset, then accuracy score will be the apt and correct parameter for evaluating the performance of each algorithm. However, in reality, the input dataset is imbalanced. It contains an unequal distribution of samples for each class. Therefore, picking an algorithm solely on its accuracy score will provide misleading results on its performance. For this reason, another metric called as the F_1 score is used. F_1 is another important parameter that can be used to test an algorithm's performance especially if it is trained on imbalanced data. Mathematically, F_1 score depends upon another set of evaluation metrics – an algorithm's precision and recall values. Precision is defined as the percentage of results that are relevant. On the contrary, recall is the percentage of total relevant results correctly predicted by the algorithm. The mathematical formulae of precision, recall and F_1 score are illustrated in Fig. 23.

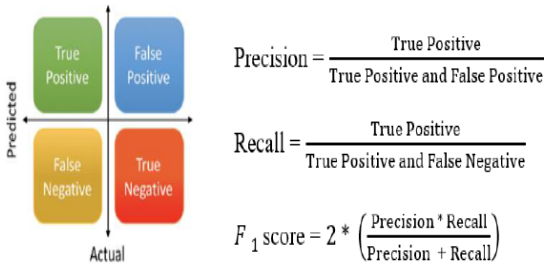


Fig. 23. Precision, Recall and F_1 score [33].

In this research paper, the dataset is multi-class, the recall and precision of each class were individually found from the confusion matrix of each algorithm. Then the average recall and precision values are calculated. Finally, the average (macro) F_1 score was calculated using these average values. The average values of the three metrics for each algorithm have been summarized with the help of Table V as shown.

From the Table V, the algorithm with the highest F_1 score is Decision Tree algorithm. Thus, decision tree algorithm has the highest performance in comparison to the other algorithms even if the accuracy score suggests otherwise. In Fig. 24, the lighting control system with the trained decision tree algorithm has been proposed.

TABLE V. Average Precision, Recall and F_1 Score of Algorithms

Algorithm	Average Precision	Average Recall	Macro F_1 score
Logistic Regression	0.8957	0.8694	0.8823
K- Nearest Neighbors	0.8873	0.8863	0.8868
Support Vector Machine	0.9017	0.88375	0.8926
Kernel SVM	0.8968	0.87359	0.8850
Naïve Bayes	0.8968	0.8735	0.8850
Decision Tress	0.8986	0.9160	0.9072
Random Forest	0.8977	0.8960	0.8969

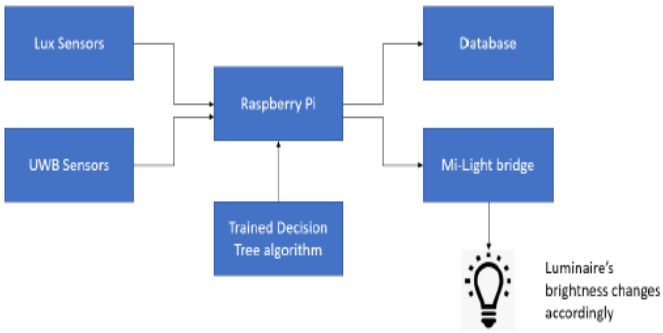


Fig. 24. Lighting Control System with the Trained Decision Tree Algorithm.

Hence, as decision tree has the best performance, it was chosen for final implementation in the lighting control system. With the help of a python program in Raspberry Pi, the trained decision tree algorithm received real time information about users' occupancy and the existing brightness in each of the zones in the room. The results were relevant output settings that were transmitted as commands to the luminaires to change their brightness accordingly. Both Fig. 25 and Fig. 26 show the change in the brightness of luminaires according to the user's position.

Fig. 25. Initial Brightness in the Room without Intelligent Lighting.

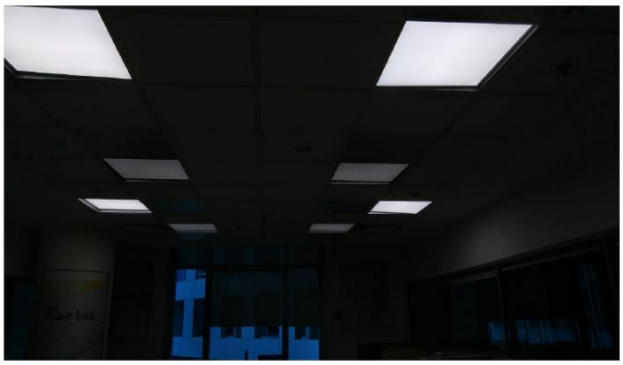


Fig. 26. Intelligent Lighting in Accordance with user's Presence and Existing Brightness in Zones 1 and 3.

VI. Conclusion

In this research paper, Lighting plays a very important role in the development, productivity, and well-being of individuals. This work focused on the development of an intelligent lighting system that can provide suitable lighting conditions in accordance with its immediate environment. It

stated the importance of good indoor lighting and also described the implementation of the system in detail. Intelligent lighting was via machine learning to suggest optimal outputs with minimum user intervention and power consumption. The proposed lighting system can be easily scaled to a wide number of applications that include lighting in homes, offices, workspaces, and buildings. As adequate lighting is always provided, it ensures user comfort, well-being, and increased security. As these lights can be accessed from any part of the world, it provides remote monitoring facility and enhances user experiences. Most importantly, as the system comprises of low -cost components that are also easily replaceable and only provide lighting when needed, it can provide huge cost and power savings.

VII. Future Scope

Another aspect that can be integrated with the existing system to increase its application range is a model that can predict the power consumption of the system. As many models to predict power consumption exist, work on the comparison between these existing models can be explored in future. This integration can greatly help in the realization of smart buildings where light and power consumption data collection methods, constant surveillance, remote monitoring facility and power savings are essential. Power consumption prediction was originally supposed to be within the current project's scope but was disrupted due to the imposed lockdown measure to fight the spread of Covid-19 pandemic. Literature review was carried out on the most extensively used algorithms for power consumption prediction. They are:

1) Support Vector Regression.
2) Artificial Neural Networks.
3) Bayesian Network.
4) Linear regression.

These algorithms can be integrated with the current system to provide useful insights on system's power consumption. One such possible approach is shown below in Fig. 27.

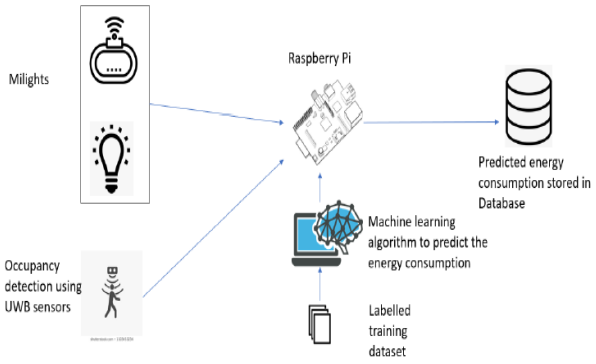


Fig. 27. Proposed System to Predict Energy Consumption.

Reference

[1] Walker MP, Liston C, Hobson JA, Stickgold R. Cognitive flexibility across the sleep-wake cycle: REM-sleep enhancement of anagram problem solving. Brain Res Cogn Brain Res. 2002.

[2] D. E. Blask. Melatonin, sleep disturbance and cancer risk. Sleep Medicine Reviews, 13:257–264, 2009.

[3] C. I. Eastman, M. A. Young, L. F. Fogg, L. Liu, and P. M. Meaden. Bright Light Treatment of Winter Depression. Archives of General Psychiatry, 55:883–889, 1998.

[4] M. Canazei. Laboratory experiment regarding impact on productivity through dynamic lighting. Technical report, Zumtobel Research, 2013.

[5] Y. Bai and Y. Ku. Automatic Light Detection and Control Using a Microprocessor and Light Sensors. IEEE Transactions on Consumer Electronics, 54(3):1173–1176, 2008.

[6] R. Magielse, S. Rao, P. Jaramillo, P. Ross, T. Ozcelebi, and O. Amft. An Interdisciplinary Approach to Designing an Adaptive Lighting Environment. In The 7th International Conference on Intelligent Environments, Nottingham, UK, July 2011.

[7] I. Chew, V. Kalavally, C. P. Tan and J. Parkkinen, "A Spectrally Tunable Smart LED Lighting SystemWith Closed-Loop Control," in IEEE Sensors Journal, vol. 16, no. 11, pp. 4452-4459, June1, 2016.

[8] S. Hong and C. Lin, "An efficient ILACS control algorithm for intelligent LED indoor lighting system," 2017 19th International Conference on Advanced Communication Technology (ICACT), Bongpyeong, pp. 29-32, 2017.

[9] I. Petrinska, V. Georgiev and D. Ivanov, "Lighting control system for public premises, based evolutionary optimization algorithm," 2018 20th International Symposium on Electrical Apparatus and Technologies (SIELA), Bourgas, pp. 1-3, 2018.

[10] Y. Gao, Y. Sun and Y. Lin, "A novel wireless lighting control strategy using RBF neural networks and genetic algorithm," 2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM), Yogyakarta, pp. 62-66, 2013.

[11] Y. Chen and Q. Sun, "Artificial intelligent control for indoor lighting basing on person number in classroom," 2013 9th Asian Control Conference (ASCC), Istanbul, pp. 1-4, 2013. doi: 10.1109/ASCC.2013.6606030.

[12] N. Khera, A. Khan, P. Biswal and C. Likhith, "Development of an Intelligent Light Intensity Control System for LED Lighting," 2018 International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, pp. 141-144, 2018. doi:10.1109/PEEIC.2018.8665595.

[13] D. Makkar and P. Syal, "Simulation of Intelligent Room Lighting Illuminance Control," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, pp. 1-4 2017. doi: 10.1109/ICCIC.2017.8524356.

[14] X. Liu and W. Wang, "Indoor Intelligent Lighting Control System Based on Power Line Carrier Design," 2010 Second WRI Global Congress on Intelligent Systems, Wuhan, pp. 408-411, 2010.doi: 10.1109/GCIS.2010.127.

[15] A. Dean and D. Voss. Design and Analysis of Experiments. Springer, 1999.

[16] Mi-light 40W 0/1~10V dimming driver PL1 User Manual, Model No. LS4, 2019.

[17] Mi-light iBox2 Wifi iBox User Manual, 2019.

[18] MDEK1001 Kit User Manual, Decawave version 1.2, 2017.

[19] DWM1xxx Product Brief, Decawave, version 1.2, 2017.

[20] DWM1001 System Overview 2.0 Manual, version 1.2 ,2018.

[21] EN-USERMAN OW-ENV-SENSOR 1.3, version 4.9.12, 2019.

[22] Raspberry Pi Model 3, raspberrypi.org/documentation.

[23] Lighting at Work, Health and Safety Executive, Second edition, published 1997.

[24] T. Mitchell. Machine Learning, chapter 8, pages 239–258. Mcgraw Hill, 1996.

[25] Pattern Recognition and Machine Learning, Page 3, 1st edition, 2006.

[26] Deep Learning (Adaptive Computation and Machine Learning series), Page 105, 2016.

[27] Page 1, Reinforcement Learning: An Introduction, 2nd edition, 2018.

[28] McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Represention" (PDF). Retrieved 22 October 2019.

[29] Tolles, Juliana; Meurer, William J. "Logistic Regression Relating Patient Characteristics to Outcomes". JAMA. 316 (5): 533–4, 2016.

[30] Cortes, Corinna; Vapnik, Vladimir N. "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362, 1995. doi:10.1007/BF00994018.

[31] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). The American Statistician. 46 (3): 175–185,1992. doi:10.1080/00031305.1992.10475879. hdl:1813/31637.

[32] Theodoridis, Sergios (2008). Pattern Recognition. Elsevier B.V. p. 203. ISBN 9780080949123.

[33] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf (2004). "A primer on kernel methods". Kernel Methods in Computational Biology.

[34] Pyle, D. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California, 1999.

Comparison of Latent Semantic Analysis and Vector Space Model for Automatic Identification of Competent Reviewers to Evaluate Papers

Yordan Kalmukov
Department of Computer Systems and Technologies
University of Ruse
Ruse, Bulgaria

***Abstract*—The assignment of reviewers to papers is one of the most important and challenging tasks in organizing scientific events. A major part of it is the correct identification of proper reviewers. This article presents a series of experiments aiming to test whether the latent semantic analysis (LSA) could be reliably used to identify competent reviewers to evaluate submitted papers. It also compares the performance of the LSA, the vector space model (VSM) and the method of explicit document description by a taxonomy of keywords, in computing accurate similarity factors between papers and reviewers. All the three methods share the same input datasets, taken from real-life conferences and the produced paper-reviewer similarities are evaluated with the same evaluation methods, allowing a fair and objective comparison between them. Experimental results show that in most cases LSA outperforms VSM and could even slightly outperform the explicit document description by a taxonomy of keywords, if the term-document matrix is composed of TF-IDF values, rather than the raw number of term occurrences.**

Keywords*—*Latent semantic analysis; vector space model; automatic assignment of reviewers to papers

I. Introduction

The assignment of reviewers to papers is probably the most important and challenging task in organizing the review process of scientific publications. Its accuracy has a direct impact on the conference/journal's quality and reputation. Submitted papers should be fairly evaluated by the most competent, in their subject domains, reviewers. To achieve that, the Program Committee (PC) chair or the assignment algorithm needs to know precisely the areas of expertise of all reviewers and the subject domains of all submitted papers. If the number of papers and reviewers is low, and all participants belong to some professional community, then it seems possible that the PC chair knows everybody, their areas of research and assigns reviewers to papers manually. However, when the number of papers and reviewers get higher, the manual assignment becomes highly inaccurate due to the lack of enough a-priori information and the many constraints (expertise, load-balancing, conflict of interests and etc.) that should be taken into account. In that case, the automatic assignment is the only accurate option. Its accuracy depends on both the assignment algorithm and the method of describing papers and reviewers' competencies. Assignment algorithms are studied in details in [1] and will not be discussed here. Instead, this article focuses on the methods of describing papers and identifying reviewers' competencies. Yes, both terms, describe and identify, are usable since methods could be explicit (users explicitly describe their papers or competencies) and implicit (subject domains and competencies are automatically identified by some piece of software).

Explicit methods usually rely on selection of keywords from a predefined list or a taxonomy [2]. They do not suffer from lack of information or sparse information, but could be a subject of incorrect, or even intentionally misleading, self-classification. Generally, choosing keywords from a predefined taxonomy of topics provides quite accurate calculation of paper-reviewer similarity factors [2].

In contrast, implicit methods do not require any additional description or actions from authors and reviewers. Instead, they rely on content analysis of both the submitted papers and the reviewers' previous publications. Implicit methods were somewhat inapplicable in the past, because reviewers whose publications cannot be found on the Internet will get their papers assigned to them at random. Currently this is not an issue anymore since all papers are published online and (at least) their abstracts are freely accessible. Fortunately, there are data aggregators such as Google Scholar, DBLP and Semantic Scholar. The latter provides an API that allows easy access to all abstracts of papers, published by a specified scientist, searching by name.

The aim of this paper is to experimentally test whether the latent semantic analysis (LSA), also known as latent semantic indexing (LSI), could be used for automatic identification of reviewers, competent to evaluate specific papers, and compare the results (in terms of accuracy) to the ones of the much simpler vector space model (VSM). The analyses are performed over real datasets taken from the CompSysTech series of conferences for a period of 5 years - from 2014 to 2018.

The paper is organized as follows: Section 2 discusses previous work from other researchers. Section 3 provides some details of how the vector space model could be used to identify competent reviewers to evaluate papers. Section 4 gives similar information but related to the use of latent semantic analysis for identifying reviewers. Section 5 describes the experimental setup and Section 6 presents the results and performs

comparative analysis between the VSM and LSA. Finally, the most important conclusions are outlined in Section 7.

II. Related Work

Commercially available conference management systems usually rely on explicit methods of describing papers and reviewers' competencies, most commonly selection of keywords/topics from a predefined list or a taxonomy [2]. However, in the recent years some of them started to implement more complex IR approaches, performing text analysis of the submitted papers and the previous reviewers' publications.

Pesenhofer et al. [3] suggest that paper-reviewer similarities are calculated as Euclidian distance between the titles of the submitted papers and the titles of all reviewers' publications. The authors evaluated their approach with data from ECDL 2005. They noted that for 10 out of 87 PC members, no publications have been found and they got their papers to review at random.

Ferilli et al. [4] use Latent Semantic Indexing (LSI, LSA) to identify reviewers to evaluate submitted papers. The document collection consisted of the titles and the abstracts of the submitted papers and the titles of reviewers' publications obtained from DBLP. Results were evaluated by the organizers of the IEA/AIE 2005 conference. In their opinion the average accuracy was 79%. According to the reviewers, the accuracy was 65% [4].

Charlin and Zemel [5],[6] propose a standalone paper assignment recommender system called "The Toronto Paper Matching System (TPMS)". It builds reviewers' profiles based on their previous publications obtained from Google Scholar or uploaded by the reviewers themselves. By using Latent Dirichlet Allocation (LDA)[1], TPMS extracts reviewers' research topics from their publications.

Dumais and Nielsen [8] used latent semantic indexing to automate the assignment of papers to reviewers in Hypertext'91 conference. Their results show a mean number of relevant articles in the top-10 of 5.9, and average precision value of 0.51. They conclude that the simple LSI method is not as good as the best human experts, but it could perform in the same general range and achieves the same performance as a human, who is not a narrow expert in the field, but has broader view and good knowledge in it [8].

Moldovan et al. [9] compare the performance of latent semantic analysis to the vector space model (VSM) applied to US patent documents from 1790 to 2005. Their results show that LSA almost always matches the VSM and sometimes slightly outperform it with an average improvement of 5%, and in a single case it performed worse with an average damage of 3% [9]. It should be noted that they were not using any term weighting model in the term-document matrix.

Many researchers (Nguyen et al. [10], Liu et al. [11], Conry et al. [12]) are proposing more complex composite methods to identify proper reviewers to evaluate papers, that also applies content analysis and IR approaches (especially LDA) on multiple data sources, not just publications' abstracts. Liu et al. [11] suggest that paper-reviewer similarities are calculated based on three aspects of the reviewer, which are lately integrated by a Random Walk with Restart (RWR) model. Authors compare their approach to other IR techniques like "text similarity" (i.e. VSM) and "topic similarity" (derived by LDA) and more or less surprisingly, their results show that text similarity actually outperforms topic similarity. So, pure VSM with proper term-weighing model could sometimes perform better than topics extraction by LDA followed by a cosine similarity of the topic vectors.

III. Using Vector Space Model (VSM) to Identify Competent Reviewers

According to the vector space model, the meaning of a document is obtained from its words. Thus, the document could be represented by an array (vector) of words. Not just its words, but all unique words from the entire document collection. This provides equal length of all document vectors and allows easy calculation of similarity between two documents by using cosine similarity. However, in case of large document collections, the vectors' length could get enormous (with most elements set to 0) that makes calculation of similarities ineffective. Fortunately, this could be overcome by using inverted index instead of forming document vectors with tens of thousands dimensions.

Document vectors do not actually contain the words (terms) themselves, but their weight instead. There are many ways of calculating term weight (called term-weighing models), but they are all based on two main components: term frequency (tf) - the number of occurrences of a term ti in the document dj; and document frequency (df) - the number of documents that contain ti. The presumption is that the more times a term occur in a document, the more important it is for that document. But, the more documents contain a term, the less informative it is. As df is an inverse measure of informativeness, we use not df, but idf – inverse document frequency. The most basic term-weighting model is the simple multiplication of tf * idf. However, there are more complex and accurate models that rely on compositions of different tf normalization functions – Singhal [13], Rousseau and Vazirgiannis [14], Robertson's BM 25 [15],[16] and others. Comparison of these models in the context of reviewer assignment problem could be found in [17]. Once terms weights are calculated, the similarity between two documents (or between the query and a document) could be easily calculated as the cosine of the angle between the two vectors.

A comprehensive experimental analysis aiming to check if the VSM could be reliably used for automatic identification of proper (competent) reviewers to evaluate papers is performed in my previous work [17]. According to the results, the short answer is "yes, it could be". It produces 5-10% less accuracy in comparison with the explicit selection of keywords from a taxonomy, but still high enough accuracy that allows the VSM to be used as a stand-alone method. Results also show that:

- The Robertson's BM 25 weighing model [15] achieves highest accuracy.
- Word stemming further increases identification accuracy.

- Using IDF only on query terms, rather than on both – the query and the documents terms, provides better results.
- Complex term-weighting models that consist of composition of different TF normalization functions provide better results than the plain lnc.ltc scheme.

Experiments were performed on real datasets, taken from the CompSysTech series of conferences for a 5 years period – from 2014 to 2018. Experiments in this study are using absolutely the same input datasets and evaluations methods, so a fair and objective comparison could be done between the vector space model and the latent semantic analysis in the context of reviewer assignment problem.

IV. Using Latent Semantic Analysis (LSA) to Identify Competent Reviewers

The latent semantic analysis (LSA) is a dimension reduction (or rank lowering) technique applied over the bag-of-words (BoW) model, that analyzes relationships between documents, but also relationships between the words they contain. The latter is very important and a major difference from the VSM. The assumption is that words which have similar meanings often occur in the same documents. Thus LSA is able to "group" semantically-related words into broader topics. The method is called "latent semantic analysis" because it discovers a number of latent (hidden) topics that could describe (separately or in combination) each document within the collection. These topics are not the exact words but they have more generalized meaning. Each word/term in the collection's dictionary is related or has a specific contribution to some topic(s). Similarly, each topic has a specific contribution to some documents. For example, the words: space, booster, shuttle, rocket, probe form the "space-related" topic. A document could be related to space if it contains any of these words. In contrast, if we apply the VSM over BoW, then each term is treated separately. For example shuttle and rocket are entirely dissimilar. However for the LSA, these terms are related. In this sense, LSA can cope with synonyms and partly with polysemy that is a great advantage in comparison to VSM. So in theory, word stemming is not necessary in preprocessing. But it will be tested during the experiments.

The input of the LSA is the term-document matrix (the leftmost part of Fig. 1) – a matrix where rows represent terms and columns represent documents. Generally, it states how many times each document contain each term, but values could be also the tf-idf weights of the terms in respect to each document.

Let's call the term-document matrix A. The row ai contains the weights of the i-th term in respect to all documents. Similarly, the row ap represents the p-th term. The dot product $a_i^T a_p$ indicates how related the i-th and the p-th terms are. Applying cosine normalization of the dot product, we get the cosine similarity between these two terms.

The matrix product AA^T will contain similarity factors between all terms in the entire document collection. Similarly, calculation of $A^T A$ provides similarities between all documents for the entire dictionary.

There is a matrix factorization technique in the linear algebra, called Singular Value Decomposition (SVD). According to it a matrix could be decomposed in three matrices such that:

$$A = U\Sigma V^T \quad (1)$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix. The values in Σ are called singular values and they show the significance of each latent topic. Values are ordered in the main diagonal in descending order, placing the most significant topic on top. The values of U are called left singular values and they indicate the contribution of each term to each discovered topic. The values of V^T are called right singular values and show the contribution of each topic to each document. The idea is illustrated with 4 terms, 3 documents (A4x3) and 2 topics on Fig. 1.

It should be noticed that in general, if A has a dimension of mxn, then the dimension of U is mxm, the dimension of Σ is mxn and dimension of V^T is nxn. However, as LSA is a dimension reduction technique, only the highest k singular values (the k most significant topics) and their corresponding singular vectors form U and V are taken into account, performing a truncated SVD. Then, as in the example above, the dimension of U is mxk, the dimension of Σ is kxk and dimension of V^T is kxn.

It could be proven that the columns of U are actually the eigenvectors of the matrix product AA^T, the columns of V (or rows of V^T) are the eigenvectors of A^TA, and the singular values of Σ are the square roots of the eigenvalues of AA^T or A^TA.

Thus, calculating SVD requires calculation of the eigenvalues and the eigenvectors of the matrix products AA^T and A^TA.

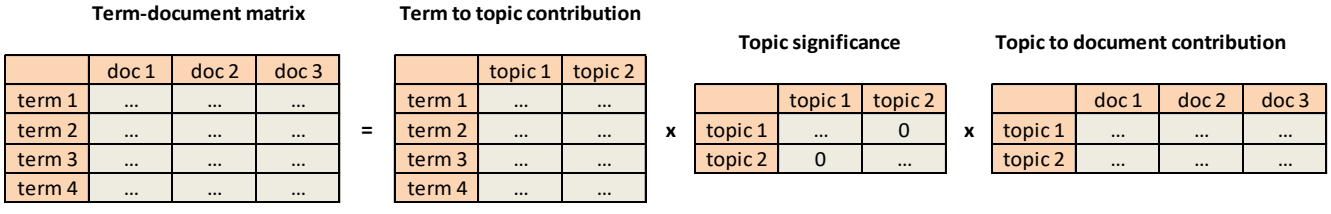


Fig. 1. SVD Decomposition of the Term-Document Matrix into Three Matrices, providing the Significance of each Latent Topic and the Contribution of Terms to Topics and Topics to Documents.

From the linear algebra, it is also known that

$$Av = \lambda v \quad (2)$$

where A is a matrix, v is an eigenvector and λ is an eigenvalue of the matrix. Equation (2) is called eigenvalue equation of A.

Equation (2) could be rewritten in the form of

$$(A - \lambda I)v = 0 \quad (3)$$

where I is the identity matrix.

Eq. (3) could have a non-zero solution (eigenvector) v, if the determinant of the matrix $(A - \lambda I)$ is zero. Thus:

$$|A - \lambda I| = 0 \quad (4)$$

So, the eigenvalues are calculated by the characteristic equation (4). The determinant of it is a polynomial function of λ with degree n, where n is the order of A. Thus the characteristic equation (4) has up to n solutions for λ which are the eigenvalues of A.

After calculating the eigenvalues of A, the eigenvector that corresponds to each eigenvalue could be determined by solving the linear equation system (3). Then the eigenvectors of AA^T form the columns of U and the eigenvectors of A^TA form the columns of V^T.

Finally, calculating the similarity between two documents in the lower dimensional k space, means calculating the cosine similarity between their corresponding columns of V^T.

If the query is missing in V^T, the original query vector should be transformed to the lower dimensional k space first (eq. 5), then the transformed query qk could be compared (by cosine similarity) with any document from V^T.

$$q_k = \Sigma_k^{-1} U_k^T q \quad (5)$$

V. Experimental Setup

Testing whether the Latent Semantic Analysis (LSA) could be successfully and reliably used to identify experts to review submitted papers is done by using real datasets taken from already conducted conferences for a period of 5 years – CompSysTech [18] from 2014 to 2018. The same datasets are used in the previous study aiming to test if the Vector Space Model (VSM) and the explicit selection of keywords from a taxonomy could be used for the same purpose. That allows completely fair and objective comparison between all these three methods of reviewer identification. All datasets contain reference values (the ground truth) for the level of competency of each reviewer in each paper he/she evaluates. These values are explicitly stated by the reviewers themselves during the review submission. So they could be considered as 100% accurate and used as a reference (benchmarks).

The document collection consists of the titles and the abstracts of all submitted papers and the titles and the abstracts of all reviewers' previous publications. The former are taken directly from the CompSysTech database, while the latter are fetched from the joint API of DBLP and Semantic Scholar. It allows getting data (full bibliography, including abstracts) from Semantic Scholar while searching by name in the DBLP's database.

Before applying the latent semantic analysis, the content of all documents is preprocessed as follows:

1) All punctuation marks (commas, dots, dashes, exclamation, quotation and question marks and etc.) are removed since they only cause troubles. If they stick to the words, that makes term recognition harder (for example "red" and "red," will be recognized as two different terms, because of the comma). If they are separated with spaces, however, they could be recognized as terms, making document vectors longer and decreasing the relative weight of meaningful terms.

2) All the text is converted to lowercase. This makes the analysis case-insensitive.

3) The text is tokenized. This is the process of splitting the text into an array (vector) of terms.

4) All semantically-insignificant terms (so called "stop words") are removed. These are prepositions, conjunctions, pronouns and etc. They are important from a syntactic point of view, but they do not represent any semantic, meaning and subject domain of the documents. Furthermore, as they are frequently appearing anywhere in the text, they will have disproportionally high tf value in comparison to the semantically-meaningful words, i.e. the semantically-insignificant stop words will highly lower the relative weight of the semantically-significant ones, which is undesirable. That's why stop words should be removed. Stop words are usually pre-defined as a list or array, and of course, they are language-dependent.

5) Finally, the Porter's word stemming algorithm [19] is applied on all remaining tokens. This is an optional step and could be skipped. Word stemming is the process of separation of word endings from the morphological root. The idea is to keep and process just the roots and skip word endings. In this way, different forms of a single word (for example: beautiful, beauty, beautifully) could be recognized as one.

The very same preprocessing is applied in the previous study [17] of the possibility of using VSM to identify reviewers. So, again, both methods are tested using the same preprocessing activities and with the same input data.

The ultimate goal of the LSA is to calculate a similarity factor between every submitted paper and every registered reviewer (PC member). It shows how competent the reviewer is to evaluate the specified paper. However reviewers have more than one publication in their profiles. Thus, a similarity factor is calculated between every submitted paper and every reviewer's publication. Then the overall similarity between a paper and a reviewer is summarized as an average of the 10% highest similarity factors between the paper and the reviewer's publications. However, the 10% number of reviewer's publications taken into account (in the overall similarity) could not be less than 3.

When performing the experiments, there are some very important settings whose value could highly impact the LSA's accuracy. These are:

- The number of latent (hidden) topics.
- The way the term-document matrix is formed. Whether it contains raw number of term occurrences or tf-idf normalized values.
- The term-weighing models in case of tf-idf normalized matrix.
- Whether word stemming is applied or not.

The number of latent topics is probably the most important setting but there is no theoretically-motivated correct value. It should be experimentally determined. Using too many topics may cause the LSA to behave like the vector space model, treating terms separately. However, choosing too few topics will cause the LSA to group unrelated terms together, losing accuracy.

For a raw term-document matrix of collection of about 5000 documents, the experiments started with 100 topics as previous research by other scientists [8], [9], [20] suggest it is a good starting point. If the number of documents and unique terms gets lower (or higher), then the number of latent topics should be decreased (increased) as well. That assumption is fully supported by the experiments in this work as well.

In general, the LSA uses a term-document matrix containing raw values, i.e. just the number of occurrences of each term in each document. However, the experiments in this article show significant increase in accuracy when the term-document matrix is composed of tf-idf term weights, rather than just the raw number of occurrences.

Experiments are performed on custom software developed in php and Matlab. The php part is responsible for extracting reviewers' publications from the Internet, building the document collection and exporting it within a proper structure in text files. The LSA is implemented in Matlab since it has a built-in function to perform the SVD decomposition.

VI. Experimental Results

To determine if the paper-reviewer similarity factors, obtained by LSA, VSM or other method are correctly calculated, they have to be compared to some reference evaluation of expertise that we trust it is correct. Since real datasets are used for experimental evaluation, fortunately, there is such a reference. During review submission, reviewers are required to explicitly indicate their level of expertise (High, Medium or Low) in respect to each paper they evaluate. As the reviewers themselves explicitly provide these levels, it could be assumed they are completely accurate and they could be used as a reference. However, the two data values (paper-reviewer similarity factors and levels of expertise) are not directly comparable. Similarity factors are decimals within the range [0.00, 1.00], while the explicitly stated levels of expertise are just "labels" – low, medium and high. To overcome this problem, a special-purpose software has been developed that converts similarity factors to levels of expertise, and then performs a correlation analysis between the automatically determined levels of expertise and the ones explicitly stated by the reviewers during the review submission. The conversion is done based on the assumption that if a reviewer ri has declared higher level of expertise than another reviewer rj (for the same paper), then ri should has higher similarity factor with the paper than rj. Detailed description of the software could be found in [21]. It is used to evaluate all the three methods – the latent semantic analysis, the vector space model and the selection of keywords from the conference taxonomy. So, again, they are all placed on equal terms (share the same input data and evaluation method) and thus could be objectively compared.

Experiments started with the data from the CompSysTech 2018 conference. Initial experiments aimed to test the influence of the previously mentioned factors – the number of latent topics, the term-weighting models and word stemming. Accuracy of the computed similarity factors is evaluated by the percentage of the correctly calculated similarities and their correlation with the levels of expertise, explicitly stated by the reviewers themselves during review submission. A similarity factor is considered to be correctly calculated, if it complies with the rules stated in [21].

The CompSysTech 2018 dataset consists of 75 submitted papers and 73 registered reviewers. After adding the abstracts of all reviewers' previous publications, the entire document collection became 4648 documents, having 21 682 unique words.

A. Experiment 1: Testing if the Number of Hidden Topics Influence the Accuracy of the Calculated Similarity Factors

The term-document matrix contains raw term frequencies, i.e. the number of occurrences of each term in each document. No stemming is applied.

As expected, results show that the number of latent topics indeed influences the accuracy of the calculated paper-reviewer similarities. Moreover, the experiment also confirm that if the document collection consists of about 5K documents and the term-document matrix contains raw tf values, then 100 is the optimal number of latent topics to start with.

B. Experiment 2: Testing if the Term-Weighting Models Influence the Accuracy of the Calculated Paper-Reviewer Similarities

In the vector space model (VSM), composite and more complex term-weighting schemas usually achieve higher accuracy than using the raw number of term occurrences. It is curious to test if this fact is valid for the LSA as well. It should be. So in this experiment, the term-document matrix does not contain the raw term frequencies (as in experiment 1), but the term weights are calculated by the basic TF-IDF model (6). Two series of experiments were performed, first with IDF applied on both the document terms and the query terms, and then with IDF applied only on query terms. TF stands for term frequency, while IDF for inverse document frequency. For more information, please refer to [17].

$$w_{i,j} = \left(1 + \log(tf_{i,j})\right) * \log(\frac{d}{df_i}) \quad (6)$$

Comparing Table I and Table II, it is clearly noticeable that the accuracy of paper-reviewer similarities gets significantly

higher when the term-document matrix is composed of TF-IDF term weights, rather than the raw number of term appearances.

TABLE I. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2018. TERM-DOCUMENT MATRIX CONTAINS RAW NUMBER OF TERM OCCURRENCES. NO STEMMING IS APPLIED

# latent topics	50	75	100	125	150
% correctly calculated	71.36 %	72.27 %	**75 %**	74.55 %	74.09 %
Pearson correlation	0.6254	0.6423	**0.6669**	0.6610	0.6588

TABLE II. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2018 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM WEIGHTING MODELS. NO STEMMING

# latent topics	10	20	30	40	50	75
Weighting model: basic TF-IDF (eq. 6), IDF applied on both document and query terms (ltc.ltc)						
% correctly calculated	79.09	79.55	**82.73**	81.36	80.91	76.82
Pearson correlation	0.7037	0.7100	**0.7610**	0.7417	0.7332	0.6819
Weighting model: basic TF-IDF (eq. 6), IDF applied only on query terms (lnc.ltc)						
% correctly calculated	75	76.82	**77.27**	74.55	74.55	75
Pearson correlation	0.6743	0.6841	**0.6867**	0.6521	0.6521	0.6513

Many researchers have proven that in VSM it is better to skip IDF for document terms and apply it just on query terms. This makes a lot of sense since a frequently appearing term in a document says it (the term) is important for the document semantics. However if IDF is applied on it, that may significantly reduce its weight, making it semantically insignificant (which is not the case). Experimental results in Table II; however, show this sense is not applicable to LSA and skipping IDF for document terms does not improve, but actually worsens accuracy.

Another interesting observation in Table II is that higher accuracy is achieved in lower number of hidden topics. This is also important since lower number of latent topics means lower dimension of the SVD transformation matrices, thus lower computational complexity and lower execution time.

C. Experiment 3: Checking if Word Stemming could Increase Accuracy

Word stemming increases accuracy in the vector space model since it recognizes different forms of a single word (for example: beautiful, beauty, beautifully) as one. However, in case of latent semantic analysis it should have minimal or no effect, because the basic idea of LSA is to group words with similar meaning together, making word stemming unnecessary. The aim of this experiment is to check this assumption.

Word stemming in this experiment is done by Poster's stemming algorithm [19] before constructing the term-document matrix. For more reliability and determination, it is tested with both the raw number of term occurrences and the TF-IDF weighting model (Table III).

TABLE III. PERCENTAGE OF CORRECTLY CALCULATED SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH REVIEWERS' OPINION FOR COMPSYSTECH 2018 WITH PORTER STEMMER APPLIED BEFORE LSA

Porter Stemmer, weighting model: raw number of term occurrences					
# latent topics	50	100	130	150	170
% correctly calculated	75	75	75.45	75.45	74.55
Pearson correlation	0.6585	0.6625	0.6614	0.6689	0.6564
Porter Stemmer, weighting model: TF-IDF (ltc.ltc)					
# latent topics	10	20	30	40	50	75
% correctly calculated	79.09	80.45	**81.82**	80	80	78.2
Pearson correlation	0.7192	0.7235	**0.7413**	0.7151	0.7213	0.70

A brief look at experiment 1 shows that without word stemming, 75% of similarity factors are correctly calculated and the correlation with reviewers' opinion is 0.6669. With a word stemming, correctly calculated similarities are 75.45% and the correlation is 0.6689. So, as expected, word stemming has an insignificant (negligible) impact on the accuracy. The combination of word stemming with TF-IDF weighted term-document matrix even lowers accuracy a little bit.

To summarize experiments 1 to 3, it can be concluded that highest accuracy (percentage of correctly calculated similarity factors and correlation with reviewers' opinion) is achieved when no word stemming is applied, and the term-document matrix is composed of TF-IDF weights, rather than raw number of term occurrences. The number of hidden (latent) topics greatly affects the accuracy as well, but it is also depends on the number of documents and unique words (terms) within the document collection, so an exact number could not be defined in advanced.

Another assumption is experimentally proven as well - that lowering the number of latent topics, increases the value of the calculated similarity factors. This is expected, but higher values of all computed similarities do not mean they are accurately calculated and real-life paper-reviewer similarities are high as well. So, the number of hidden topics should not be lowered too much or it may highly distort the results. Experiments show that in case of TF-IDF weighted term-document matrix, going down to 5 or less topics, produces very high values (>0.9) for all paper-reviewer similarities, which of course cannot be true.

D. Experiment 4: Testing LSA with other CompSysTech Datasets

To verify that results for CompSysTech 2018 are not obtained by a lucky chance, the latent semantic analysis is applied (without word stemming) on all CompSysTech issues for a 5 year period of time – from 2014 to 2018.

It should be noted here that when downloading the abstracts of reviewers' previous publications, only manuscripts published before the specific conference year are taken into account. For example, if the conference is in 2015, then

reviewers' publications up to 2014 (including) are considered for processing. For that reason the document collection of CompSysTech 2014 will be smaller than the one of 2018, regardless of the number of actual reviewers.

The percentage of correctly calculated paper-reviewer similarities and the level of their correlation with the reviewers' opinions for the other CompSysTech issues (2017 to 2014) are presented in Tables IV to VII. As expected, the highest accuracy (marked in green) in all cases is obtained with TF-IDF weighted term-document matrix. However, it could be seen that going back in time, it is achieved at lower number of latent topics – 30 for CompSysTech 2018, and just 15 for CompSysTech 2014, 2015 and 2016. That is expected and pretty logical – as we go back in time, the document collection gets smaller (from 4648 to 2550 documents) due to the lower number of reviewers' publications. The smaller the document collection, the lower is the number of unique words, leading to lower optimal number of hidden topics.

TABLE IV. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2017 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 107, Reviewers: 76 Documents: 4128, Unique words: 19698					
Weighting model: raw term frequencies					
# latent topics	80	100	120	140	
% correctly calculated	76.19	76.83	77.46	75.56	
Pearson correlation	0.7482	0.7579	0.7647	0.7460	
Weighting model: basic TF-IDF (eq. 6), ltc.ltc					
# latent topics	10	20	25	30	40
% correctly calculated	80	81.59	**81.9**	80.95	80.63
Pearson correlation	0.7867	0.8019	**0.8078**	0.7980	0.7916

TABLE V. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2016 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 117, Reviewers: 73 Documents: 3926, Unique words: 19787				
Weighting model: raw term frequencies				
# latent topics	60	80	100	120
% correctly calculated	73.93 %	74.79 %	74.5 %	73.93 %
Pearson correlation	0.6956	0.7056	0.6974	0.6876
Weighting model: basic TF-IDF (eq. 6), ltc.ltc				
# latent topics	15	20	25	30
% correctly calculated	**80.8 %**	80.52 %	79.66 %	78.22 %
Pearson correlation	**0.7500**	0.7451	0.7414	0.7250

TABLE VI. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2015 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 103, Reviewers: 74 Documents: 3090, Unique words: 17165				
Weighting model: raw term frequencies				
# latent topics	40	50	60	80
% correctly calculated	76.95 %	76.95 %	76.27 %	75.25 %
Pearson correlation	0.7280	0.7383	0.7388	0.7213
Weighting model: basic TF-IDF (eq. 6), ltc.ltc				
# latent topics	8	10	15	20
% correctly calculated	81.69 %	82.71 %	**82.71 %**	81.36 %
Pearson correlation	0.7720	0.7842	**0.7874**	0.7717

TABLE VII. PERCENTAGE OF CORRECTLY CALCULATED PAPER-REVIEWER SIMILARITIES AND LEVEL OF THEIR CORRELATION WITH THE EXPLICIT REVIEWERS' OPINION FOR COMPSYSTECH 2014 AT DIFFERENT NUMBER OF LATENT TOPICS AND TERM-DOCUMENT MATRIX'S WEIGHTING MODELS

Papers: 107, Reviewers: 65 Documents: 2550, Unique words: 14810				
Weighting model: raw term frequencies				
# latent topics	30	40	60	80
% correctly calculated	74.1 %	74.75 %	73.77 %	71.8 %
Pearson correlation	0.6791	0.6835	0.6756	0.6574
Weighting model: basic TF-IDF (eq. 6), ltc.ltc				
# latent topics	8	10	15	20
% correctly calculated	78.69 %	78.69 %	**79.67 %**	78.36 %
Pearson correlation	0.7114	0.7074	**0.7340**	0.7137

Finally, it is interesting to see a direct performance comparison between the latent semantic analysis (LSA), the vector space model (VSM) and the explicit document description by a taxonomy of keywords in computing paper-reviewer similarities for all issues of CompSysTech. Such a comparison is presented in Table VIII. It includes only the highest accuracies, obtained by the VSM and LSA for every CompSysTech issue. Data for the VSM and the method of describing papers/reviewers by taxonomy of keywords are taken from my previous publication [17]. All methods are tested by using the same input data (CompSysTech 2014-2018 datasets) and by the same similarity factors' evaluation tool [21]. So, comparison is fair and objective.

There are dozens of experiments, testing many popular TF-IDF weighting models with the VSM in [17]. However, Table VIII shows only the best performing one – the algebraic version of Robertson's BM 25.

Expectedly, LSA outperforms VSM, even for the best performing term-weighting model for VSM. However, it is a bit surprising that, in some cases, LSA slightly outperforms the explicit document description by taxonomy of keywords as well.

TABLE VIII. COMPARISON OF LSA, VSM AND THE EXPLICIT DOCUMENT DESCRIPTION BY A TAXONOMY OF KEYWORDS FOR ALL COMPSYSTECH ISSUES FROM 2014 TO 2018

	CST 2018		**CST 2017**		**CST 2016**		**CST 2015***		**CST 2014***	
Total assignments	220		315		349		295		305	
	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation	Correctly calculated, %	Correlation
Taxonomy of keywords [2]										
	81.82	0.75	81.90	0.80	80.23	0.74	**85.08**	**0.81**	**80.66**	**0.78**
Vector Space Model (VSM), Robertson's BM25 TF_{kep} / TF_{kep} x IDF										
No stemming	73.64	0.67	76.51	0.74	72.78	0.65	75.59	0.70	72.13	0.67
Porter stemmer	74.09	0.66	79.37	0.77	73.64	0.68	76.95	0.72	74.43	0.68
Latent Semantic Analysis (LSA), TF-IDF weighted term-document matrix, eq. (6)										
No stemming	**82.73**	**0.76**	**81.90**	**0.81**	**80.80**	**0.75**	82.71	0.79	79.67	0.73

* Three PC members of CompSysTech 2015 and 2014 were not identifiable in DBLP.

It should be noted here, that 3 PC members of CompSysTech 2014 and 2015 were not found in DBLP, so the abstracts of their previous publications were excluded from the document collection, meaning they get zero similarities with all papers. Actually, missing data for some reviewers is the highest threat to LSA and VSM since they calculate similarities based on content analysis. If there is no content, there is no similarity, and those reviewers could have their papers assigned at random.

Results of the LSA to VSM comparison comply with most of the previous similar research. Although the LSA achieves an increase of 30% in the average accuracy for the MED collection, it shows much lower improvement for CISI and NPL datasets, while performing even worse for TIME and CACM collections [22]. In real-life applications, improvement is also moderate. Moldovan et al. [9] applied both LSA and VSM to analyze US patent documents and their results show that LSA slightly outperform VSM with an average improvement of up to 5%. That is fully comparable to the results obtained in this study, in case the term-document matrix is composed of raw term frequencies. However, if the term-document matrix is composed of tf-idf weights, accuracy could be increased with up to 10% in respect to the VSM.

VII. CONCLUSION

After performing large number of experiments with all the five CompSysTech datasets, it can be concluded that:

1) The latent semantic analysis (LSA) could be accurately and reliably used to identify competent reviewers to evaluate papers.

2) The latent semantic analysis outperforms the vectors space model in almost all cases, even when VSM implies the Robertson's BM 25 as a term-weighting model.

3) When the term-document matrix of LSA is composed of raw number of term occurrences, the LSA slightly outperforms VSM by 2-3 ppts (percentage points).

4) Composing the term-document matrix of TF-IDF weights, rather than raw number of term occurrences, additionally boosts accuracy by further 5 ppts, and allows the LSA even to slightly outperform the method of explicit document description by a taxonomy of keywords.

5) In contrast to the vector space model, the LSA achieves higher accuracy when IDF is applied to both document and query terms.

6) Word stemming has a little effect on accuracy of similarities computed by LSA.

7) The optimal number of latent (hidden) topics depends on the number of unique words (terms) within the document collection. Higher number of terms results in higher optimal number of latent topics, and the opposite.

8) Lowering the number of latent topics increases the values of all calculated paper-reviewer similarities, but not their accuracy.

9) The highest threat in using LSA to assign reviewers to papers is to have a PC member who cannot be found in DBLP and Semantic Scholar. In this case, no publications could be extracted for him/her and he/she will get zero similarities with all papers. The latter means that papers will be assigned to him/her at random.

Both the latent semantic analysis and the vector space model could be reliably used to identify competent reviewers to evaluate papers. LSA achieves higher accuracy, but it is harder to be implemented and has higher time complexity. Furthermore, in contrast to the VSM, the LSA could not be computed by using an inverted index, making it much slower than VSM. Additionally, the accuracy of LSA depends on the number of latent topics, but the optimal number could not be

set in advanced. So, although LSA achieves higher accuracy, the VSM may be a better choice for commercially available conference management systems due to its simplicity and better time complexity, allowing real-time computation even for large scale conferences.

Other IR approaches (most probably composition of several methods and/or data sources) will be tested in future to check if they could also be used to identify competent reviewers to evaluate submitted papers. So far, both the VSM and the LSA, together with the method of explicit description of papers and reviewers by choosing keywords from a predefined taxonomy, turned to be quite reliable option for this task.

REFERENCES

[1] Y. Kalmukov, "An algorithm for automatic assignment of reviewers to papers", Scientometrics, 2020, No 124 (3), pp. 1811–1850, https://doi.org/10.1007/s11192-020-03519-0.

[2] Y. Kalmukov, "Describing Papers and Reviewers' Competences by Taxonomy of Keywords", Computer Science and Information Systems, 2012, No 9(2), pp. 763-789, https://doi.org/10.2298/CSIS110906012K.

[3] A. Pesenhofer, R. Mayer, A. Rauber, "Improving Scientific Conferences by enhancing Conference Management System with information mining capabilities", Proceedings IEEE International Conference on Digital Information Management (ICDIM 2006), ISBN: 1-4244-0682-x; S. 359 - 366.

[4] S. Ferilli, N. Di Mauro, T.M.A. Basile, F. Esposito, M. Biba, "Automatic Topics Identification for Reviewer Assignment", 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006. Springer LNCS, 2006, pp. 721-730.

[5] L. Charlin and R. Zemel, "The Toronto paper matching system: an automated paper-reviewer assignment system." (2013).

[6] L. Charlin, R. Zemel and C. Boutilier, "A framework for optimizing paper matching", In Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (Corvallis, OR, 2011). AUAI Press, 86–95.

[7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research 2003, 3:993-1022.

[8] Susan T. Dumais, and Jakob Nielsen, "Automating the assignment of submitted manuscripts to reviewers." In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 233-244. 1992.

[9] Andreea Moldovan, Radu Ioan Bot and Gert Wanka, "Latent semantic indexing for patent documents." (2005).

[10] Jennifer Nguyen, Germán Sánchez-Hernández, Núria Agell, Xari Rovira, and Cecilio Angulo, "A decision support tool using Order Weighted Averaging for conference review assignment", Pattern Recognition Letters 105 (2018): 114-120.

[11] Xiang Liu, Torsten Suel and Nasir Memon. "A robust model for paper reviewer assignment", In Proceedings of the 8th ACM Conference on Recommender systems, pp. 25-32. 2014.

[12] Don Conry, Yehuda Koren, and Naren Ramakrishnan. "Recommender systems for the conference paper assignment problem", In Proceedings of the third ACM conference on Recommender systems, pp. 357-360. 2009.

[13] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization", In Proceedings of SIGIR'96, pages 21–29, 1996.

[14] F. Rousseau and M. Vazirgiannis, "Composition of TF normalizations: new insights on scoring functions for ad hoc IR", In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 917-920. 2013.

[15] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of documentation, vol. 60 no. 5, pp 503–520, 2004.

[16] S. E. Robertson, S. Walker, K. Spärck Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3", In Proceedings of TREC-3, pages 109–126, 1994.

[17] Y. Kalmukov, "Automatic Assignment of Reviewers to Papers Based on Vector Space Text Analysis Model", Proceedings of the 21st International Conference on Computer Systems and Technologies, CompSysTech 2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 229–235, DOI: https://doi.org/10.1145/3407982.3408026.

[18] CompSysTech – International conference on computer systems and technologies, http://www.compsystech.org.

[19] Matrin F. Porter, "An algorithm for suffix stripping", In J. S. Karen and P. Willet, editors, Readings in information retrieval, pages 313-316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[20] D. Grossman, O. Frieder, "Information Retrieval: Algorithms and Heuristics", 2nd Ed. Springer, The Netherlands, 2004, ISBN: 1-4020-3004-5.

[21] Y. Kalmukov, "A Software Tool for Accuracy Evaluation of Calculated Similarity Factors between Papers and Reviewers", 2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE), 2020, pp. 1-5, IEEE, https://doi.org/10.1109/EEAE49144.2020.9279032.

[22] Dandan Li and Chung - Ping Kwong, "Understanding latent semantic indexing: A topological structure analysis using Q - analysis", Journal of the american society for information science and technology 61, no. 3 (2010): 592-608.

Evaluation of Applicability of 1D-CNN and LSTM to Predict Horizontal Displacement of Retaining Wall According to Excavation Work

Seunghwan Seo, Moonkyung Chung*
Department of Geotechnical Engineering Research
Korea Institute of Civil Engineering and Building Technology, Goyang-si, Republic of Korea

***Abstract*—During excavation works in downtown, stability and safety considerations of such excavations and constructions are crucial for which continuous wall structures with varying structural components are commonly used. Most of the current models used for this purpose are often complex, where the accepted parameters do not have a clear physical meaning. Moreover, accurate ground movement forecasts are challenging due to nonlinear and inelastic soil behavior. Therefore, this study proposes a method to predict the lateral displacement of the braced wall at each stage of excavation by using all the basic information necessary for braced wall design, including ground information of the excavation site, support methods such as the type of brace, location and stiffness, information about the neighboring buildings, and the results of numerical analysis. One-dimensional convolutional neural network and long short-term memory network are used for estimation and prediction to develop an optimal prediction model based on well-refined but limited data. The applicability of the braced wall was confirmed for safety management by predicting the horizontal displacement of the braced wall for each stage of excavation. The proposed model can be used to predict the stability of the horizontal wall for each excavation step and reduce accident risks, such as collapse of the retaining wall, which may occur during construction.**

***Keywords*—Excavation; wall displacement; neural network; prediction wall deflection; CNN-LSTM**

I. Introduction

Owing to the increase in high-rise buildings in urban areas, an increasing number of excavations are being planned. It is important to consider the potential serviceability issues caused by the construction of these structures. To maintain the stability of excavations, continuous wall structures with varying structural components are commonly used. These components can help prevent ground movements and reduce the impact on nearby structures. A reasonable estimate of the lateral wall deflection profiles caused by braced excavations is critical to ensure safe and economical construction. Therefore, measurements during construction are critical for determining the stability of the braced wall during excavation. In particular, wall displacement is the primary sign of problems with stability of the braced wall. To monitor this during construction, it is periodically measured using an inclinometer that can measure the lateral displacement of the braced wall, and thus the risk is determined. In addition, the inclinometer is the only method for measurement of the lateral displacement of the braced wall during excavation throughout the entire construction stage. Therefore, monitoring the lateral wall displacement through the inclinometer is indispensable; however, the measurement cost increases drastically if it is installed on all braced walls at an excavation site. Currently, engineers measure the lateral displacement of the braced wall using an inclinometer at a section that is representative of the entire structure. Therefore, there are still limits to management in sections other than the representative section, and accidents sometimes occur in these sections. As such, the retaining wall displacement for the unmeasured section can be estimated using numerical analysis or interpolation of the database. However, it is difficult to accurately predict ground movement because soil is a complex material and has inelastic behavior. Although various numerical models consider various features of soil, many of these models are often complex, and the accepted parameters do not have a clear physical meaning. Factors that affect the behavior of retaining walls at excavation sites are very diverse, such as the type of ground, the presence of adjacent buildings, and the support and wall construction methods. Based on empirical analysis of measured displacements in a large number of case histories, it is a proven method [1-4] to identify the main parameters affecting the deformation behavior during excavation works, as well as to examine general trends and patterns. This empirical design method is currently used a lot by engineers, but it is more inaccurate than a numerical model. However, it requires enormous computing resources to use a numerical model to predict the retaining wall. Therefore, an artificial intelligence (AI) based approach in geotechnical engineering is being used to analyze the complex behavior of underground structures.

An artificial neural network (ANN) was used in many research [5-16] to estimate the lateral wall displacement in excavation works. As some research trend, ANN was also used by Kung et al. [11] to calculate the deflection of diaphragm walls caused by excavation in clays. Chern et al. [12] used a back-propagation neural network (BPNN) model to forecast lateral wall displacement in top-down excavation. Random forest (RF) algorithm was utilized by Zhou et al. [13] to anticipate ground settlements caused by the building of a shield-driven tunnel. For the inverse analysis of soil and wall parameters in braced excavation, Zhang et al. [14] used multivariate adaptive regression splines (MARS). For the determination of Earth Pressure Balance (EPB) tunnel-related maximum surface settlement, Goh et al. [15] used the MARS

*Corresponding Author.

model. Xie and Peng [16] tested the prediction power of Random Forest (RF) modeling for estimating tunnel Excavation Damaged Zones (EDZs). Despite the widespread application of supervised learning algorithms in geotechnical engineering, they have not been frequently applied for lateral wall displacement prediction in deep braced excavations considering the anisotropic shear strength.

As such, various artificial intelligence techniques have been utilized in relation to the stability of the retaining wall at the excavation site. However, the research so far has been limited to the study of the prediction of the maximum displacement of the retaining wall at the time when the excavation work is completed. In order to determine the stability during excavation work, it is important to manage the displacement of the retaining wall during construction, that is, according to the excavation stage. Most of the accidents related to excavation work occur during the excavation process, but no attempt has been made to predict the displacement of the retaining wall during excavation work. Therefore, predicting not only the maximum displacement after the excavation work is completed, but also the displacement of the retaining wall at each stage of excavation is considered to be helpful in reducing collapse accidents that occur in actual excavation work and evaluating the stability of the retaining wall.

This study attempted to predict the lateral displacement of the braced wall at each stage of excavation by using all the basic information necessary for braced wall design, including ground information of the excavation site, support methods such as the type of brace, location, and stiffness, information about the neighboring buildings, and the results of numerical analysis. Therefore, one-dimensional convolutional neural network (1D-CNN) and long short-term memory (LSTM) network were used, and the applicability of the braced wall was confirmed for safety management by predicting the horizontal displacement of the braced wall for each stage of excavation.

II. Prediction Model and Construction Methodolgy

A. 1-D CNN

Predictions based on existing time series data mainly use deep learning algorithms [17, 18]. CNN (Convolutional Neural Network) is a deep learning algorithm and an effective neural network for identifying patterns in data because it specializes in processing array data. Therefore, CNN utilizes various filters that can be used as shared parameters; in the case of two dimensions, it efficiently extracts and learns features from adjacent images while maintaining the spatial information of the image. CNN, which mainly uses two-dimensional data, can be applied to data feature extraction and data prediction analysis by utilizing one-dimensional time series data [19-21]. CNN has the advantage of enabling easier training based on minimal parameters and preprocessing of data. The following equation (1) describes the output of a CNN corresponding to one-dimensional input data.

$$s(t) = (x * w)(t) = \sum x(a)w(t-a) \quad (1)$$

where x is the input data, w is the kernel map, and $s(t)$ is the feature map, which is the output layer. The CNN algorithm consists of four steps. In the first step, the kernel, which has a weighted function as the input data, traverses in a certain flow, and several convolution products are calculated in parallel. In the second step, the values computed in parallel go through the activation function, and the features of the input data are detected and output to the feature map. In the third step, the pooling function is used in the pooling layer to reduce the feature data detected in the feature map. As described above, the CNN algorithm extracts the features of the data through the iterations of the CNN and pooling layers. In the last step, for the dataset extracted from the CNN and pooling layers, the data constructed in an array are transformed into a column vector array through the fully connected layer, and the features of the data are classified. Fig. 1 shows the structure of the 1D-CNN algorithm.

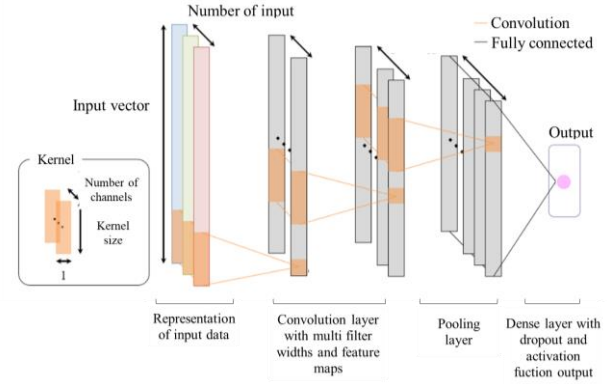


Fig. 1. Structure of 1-D CNN Algorithm.

B. LSTM

LSTM(Long-Short Term Memory) is mainly used for prediction and classification studies such as genes, handwriting, voice signals, sensor data, and stock prices [22]. Recently, many studies have been conducted to improve the prediction performance by modifying the structure of the LSTM [23, 24]. The LSTM algorithm was developed to solve the problem that owing to the structure of the recurrent neural network (RNN) algorithm, the time-series data of the distant past are not reflected if the data are large. The RNN algorithm transforms the hidden layer into forget, input gate, and output gates, which controls the flow of information to reflect time-series data of the distant past. Fig. 2 shows the structure of the LSTM algorithm [25], in which X represents the input layer, h represents the output layer, and a represents the hidden layer transformed into forget, input, and output gates.

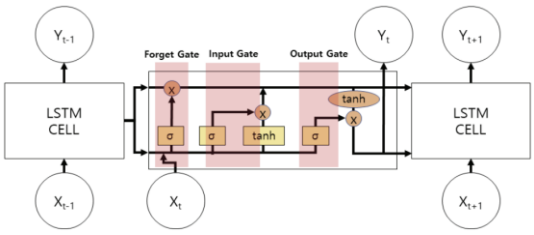


Fig. 2. Structure of LSTM.

C. Proposed Model

The 1D CNN-LSTM model proposed in this study is a retaining wall displacement prediction algorithm to produce an optimal learning effect with limited iterative learning of time-series data by combining CNN and LSTM (Fig. 3). The structure of the 1D CNN-LSTM model is divided into three stages. The first stage has a three-layer CNN structure and max pooling. In the first stage, the periodic and non-periodic features of the time-series data are extracted from the CNN layer, and a feature map is created using the output values. The max pooling layer is used to reduce the size of the extracted feature data. Max pooling selects the maximum value of the feature map. This process was repeated three times to extract the periodic and non-periodic features of the time-series data, and the data size was reduced significantly compared to the initial data size. The second stage consists of a flattened layer and a dense layer. The flattened layer converts multi-dimensional array data into 1D time-series data, and the dense layer connects both inputs and outputs. In the third stage, deep iterative learning of the LSTM layer was performed to ensure that the LSTM layer learns the relationship between the past and future data through the CNN. Future data were predicted based on the learned relationship.

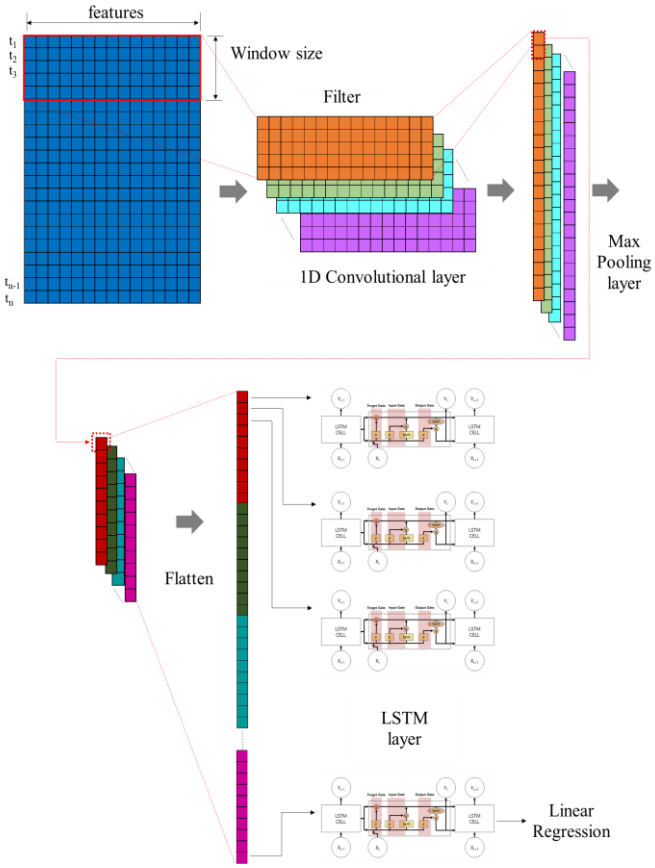


Fig. 3. Structure of Proposed Wall Deflection Prediction Model.

D. Data Collection and Preparation

Data collection is one of the most crucial steps in the prediction modeling. In this study, we need training data for the design and measurement over time to predict the horizontal displacement of the retaining wall for each excavation step. Therefore, we used the data obtained from excavation work sites in South Korea to prepare 30 input datasets by sorting the soil information, member information of the temporary retaining wall, numerical analysis results, and measurement results for each excavation step. The variables of the retaining wall data included all factors affecting the displacement of the retaining wall, such as the location, ground layer formation, soil strength, height of the retaining wall, height of the upper weak layer, retaining wall type, rigidity of the retaining wall, support type, and horizontal displacement of the ground. Fig. 4 is an example in which the various variables used as input data are scaled to a value between 0 and 1 and organized by depth. We could not collect a large amount of data because it was difficult to collect relevant information for step-by-step prediction from actual excavation sites. Therefore, the number of training data used in this study was relatively small, and we attempted to find the optimal model through cross-validation by changing the training and validation data.

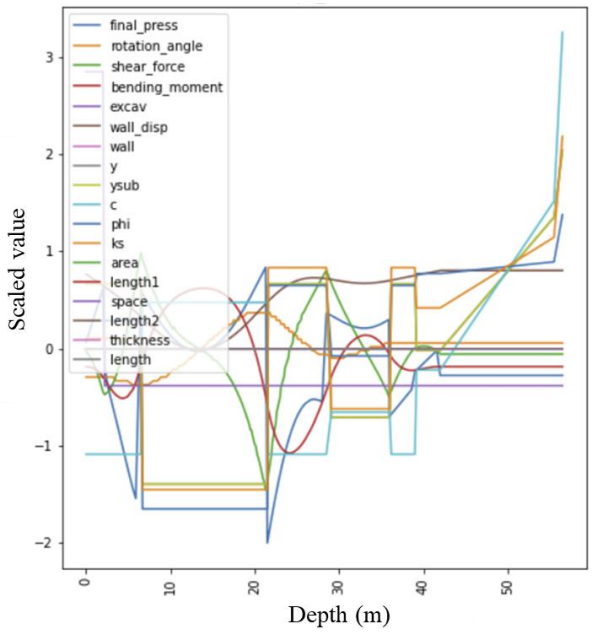


Fig. 4. Example of Preprocessed Input Data.

III. Results and Discussion

Fig. 5 shows the overall accuracy and loss of the training and validation datasets. This shows that both the training loss and the validation loss start to converge above the 100th epoch. During this time, the overall accuracy of the training and validation tends to remain stable. Finally, the training was conducted for 1,000 epochs, and the optimal result was obtained at the 210th epoch. The performance improvement of the model cannot be expected through further training.

Because it was difficult to collect all excavation data for each step of the excavation work, the prediction values through cross-validation in this study were validated in this study. After training the model by excluding the design values of certain excavation site locations, the model by comparing the prediction values to the design values of those site locations were validated.

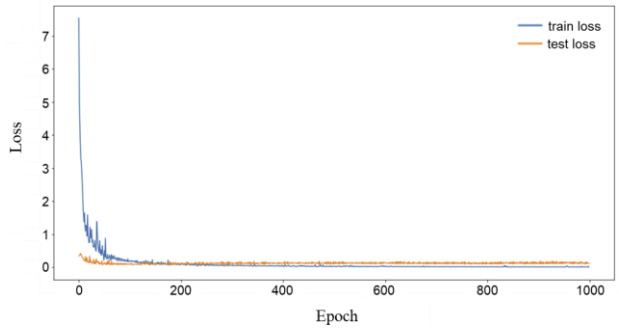


Fig. 5. Loss per Epoch during the Training of the Predicting Wall Deflection.

Fig. 6(a) and (b) show the predictions of the design values of excavation sites A and B, respectively, for each of the three excavation steps. The prediction results show that the changing trend of the horizontal displacement of the retaining wall is predicted properly in most cases. However, the horizontal displacement of the retaining wall was overestimated or underestimated at certain sites because the soil conditions and construction methods of the excavation sites were all different.

Fig. 6(a) shows that for site A, the horizontal displacement trend of the retaining wall is predicted well in every excavation step. Although there is a tendency to slightly overestimate the maximum horizontal displacement compared to the design value (true value), the depth at which the maximum horizontal displacement occurs matches exactly in each excavation step. This could help predict the position at which risk occurs before construction. Furthermore, the prediction values were mostly similar to the true values, regardless of the excavation depth in each excavation step.

In Fig. 6(b), site B also shows that the horizontal displacement trend of the retaining wall is predicted well in every excavation step. Furthermore, the prediction value matched the true value for the depth at which the maximum horizontal displacement occurred. However, in contrast to site A, the maximum horizontal displacement of the retaining wall was underestimated. In every case, it was determined that the accuracy of the prediction increases as the excavation progresses, and if the amount of training data increases, higher accuracy can be expected.

Fig. 7(a) and (b) show the predictions of the horizontal displacement of the retaining wall for certain cross-sections of sites A and B, respectively, for each excavation step. Here, the true value refers to the value measured using an inclinometer. For site A, it can be seen that the inclinometer measurement value and the prediction value match well in each excavation step. Furthermore, the predicted maximum horizontal displacement of the retaining wall is almost the same as the actual measurement value, and as the excavation progresses step-by-step, the difference from the actual measurement value decreases. For site B, few errors appeared to occur at low depths, but the trend of the displacement profile of the retaining wall was consistent. In actual measurements, the traffic on the surrounding roads and the adjacent buildings affect the ground. However, it is difficult to prepare these values in detail in the training data. Therefore, errors occurred at low depths close to the ground surface.

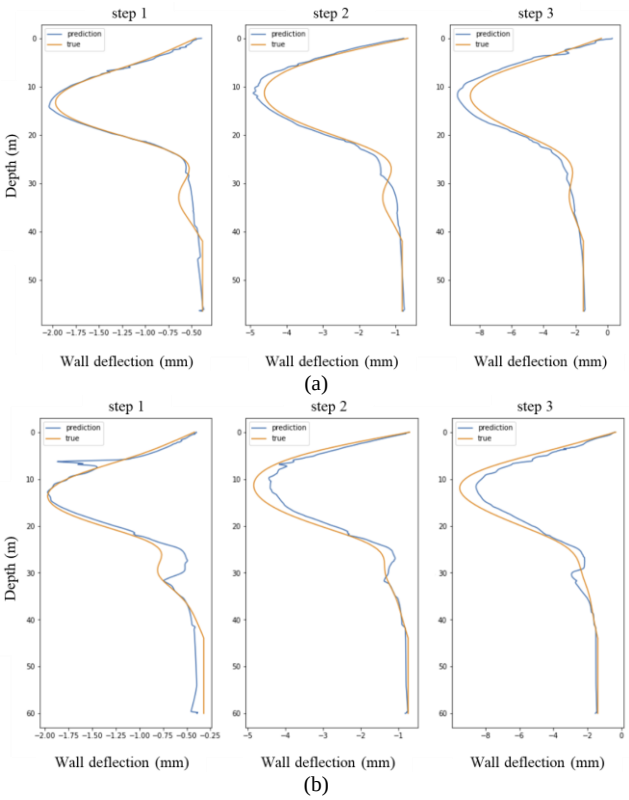


Fig. 6. Wall Deflection Profiles with Numerical Analysis versus Prediction (a) Excavation Site A, (b) Excavation Site B.

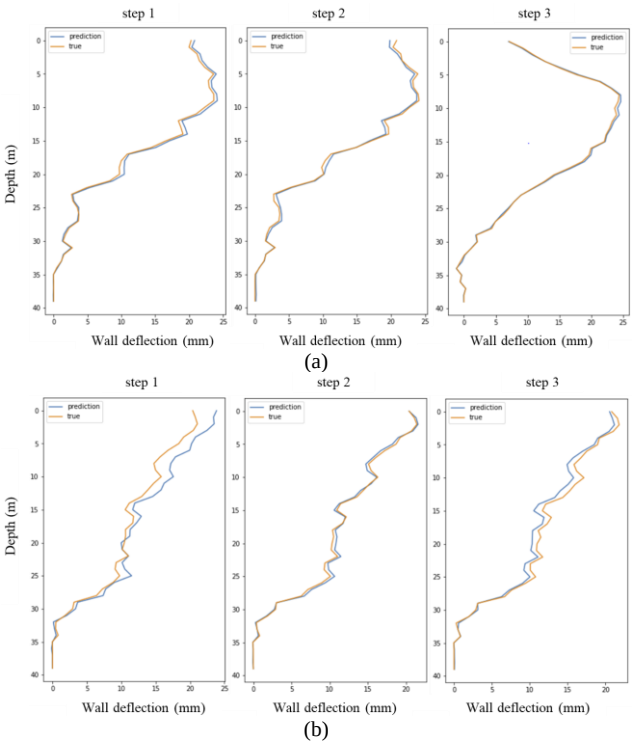


Fig. 7. Measured Wall Displacement Profiles versus Prediction (a) Excavation Site A, (b) Excavation Site B.

IV. Discussion

In this study, a method combining CNN and LSTM was applied to predict the displacement of the retaining wall according to the excavation work step by step, and its applicability was confirmed. Through this study, it was found that the displacement change of retaining wall can be predicted as the excavation work progresses. While previous studies were limited to predicting the maximum displacement of the retaining wall, this study made it possible to measure the entire profile of the retaining wall. The ground inclinometer is the only measurement item that can reflect the entire excavation process, and the prediction accuracy can be improved by using this measurement value. In the previous study [9, 10, 14] the maximum displacement prediction error of the retaining wall was about 6~23%, but in this study, the prediction error for the maximum displacement was about 3~18%. In addition, the prediction error for each stage of excavation was also confirmed to be about 15%. Since the machine learning materials used so far may not be able to represent various environments such as all ground conditions and retaining wall construction methods, prediction errors may appear differently depending on the characteristics of the site. However, it is judged that the signal of accidents can be confirmed in advance by comparing the predicted results using the model proposed in this study with the retaining wall management standard according to the excavation work.

V. Conclusion

This study proposes a model that predicts the retaining wall displacement for each excavation step by combining 1D CNN and LSTM using the retaining wall measurement data. Compared to previous studies on the prediction of the maximum displacement of the retaining wall, this study has the advantage that the displacement profile of the retaining wall can be predicted for each excavation step. For highly reliable predictions, we need a large amount of data; however, we aimed to propose an optimal prediction model based on well-refined data by training the model using limited training data and combining 1D CNN and LSTM. The results predicted by applying the measurement data of actual sites in the proposed model showed few differences from the actual measurement values. In these results, there was a tendency to show errors at certain sites because each site has different soil conditions and construction methods. Furthermore, this phenomenon seemed to be caused by the limited number of data, and this problem is expected to be solved by inputting additional measurement data in future.

For the safe management of the retaining wall during excavation work, predictions are required for not only the measurements of representative cross-sections but also for the unmeasured sections. In this regard, the proposed prediction model of this study can be used to predict the stability of the retaining wall for each excavation step and reduce accident risks, such as collapse of the retaining wall, which may occur during construction. Although the proposed model has some limitations, if appropriate data for the proposed model are collected and the database is built upon them, it could potentially help experts to use the model for designing or constructing retaining walls. Furthermore, it can help perform a more economical and safer retaining wall design or construction.

Acknowledgment

This research was supported by a grant from the project “Development of Smart Complex Solution for Large-Deep Underground Space Using Artificial Intelligence”, which was funded by the Korea Institute of Civil Engineering and Building Technology(KICT).

References

[1] Long M, “Database for retaining wall and ground movements due to deep excavations,” J. Geotech. Geoenviron. Eng., 127, pp. 203-224, 2001.

[2] Moormann C, “Analysis of wall and ground movements due to deep excavations in soft soil based on a new worldwide database,” Soils Found. Jpn. Geotech. Soc. 44(1), pp. 87-98, 2004.

[3] Wang ZW, Ng CWW, Liu GB, “Characteristics of wall deflections and ground surface settlements in Shanghai,” Can. Geotech. J., 42, pp. 1243-1254, 2005.

[4] Ali J, Khan AQ, “Behaviour of anchored pile wall excavations in clays,” Geotechn. Eng. 170(6), pp. 493-502, 2017.

[5] Goh, A.T.C., Wong, K.S., Broms, B.B., “Estimation of lateral wall movements in braced excavations using neural networks. Can. Geotech. J. 32 (6) (1995) 1059-1064.

[6] Hsiao, E.C.L., Kung, G.T.C., Juang, C.H., Schuster, M., Estimation of wall deflection in deep excavation – neural network approach,” Geoshanghai Int. Conf., GSP, 155, pp. 348-354, 2006.

[7] Goh, A.T.C., Zhang, W.G., “An improvement to MLR model for predicting lique- faction-induced lateral spread using Multivariate Adaptive Regression Splines,” Eng. Geol., 170, pp. 1-10, 2014.

[8] Adoko, A.C., Jiao, Y.Y., Wu, L., Wang, H., Wang, Z.H., “Predicting tunnel con- vergence using multivariate adaptive regression spline and artificial neural network,” Tunn. Under. Space Technol. 38(3), pp. 368-376, 2013.

[9] Zhang, W.G., Goh, A.T.C., Xuan, F., “A simple prediction model for wall deflection caused by braced excavation in clays,” Comput. Geotech., 63 pp. 67-72, 2015.

[10] Zhang, W.G., Goh, A.T.C., “Multivariate adaptive regression splines and neural network models for prediction of pile drivability,” Geosci. Front. 7, pp. 45-52, 2016.

[11] Kung, G.T.C., Hslao, E.C.L., Schuster, M., Juang, C.H., “A neural network approach to estimating deflection of diaphragm walls caused by excavation in clays,” Comput. Geotech. 34 (5), pp. 385-396, 2007.

[12] Chern, S., Tsai, J.H., Chien, L.K., Huang, C.Y., “Predicting lateral wall deflection in top–down excavation by neural network,” Int. J. Offshore Polar Eng., 19 (2), pp. 151-157, 2009.

[13] Zhou, J., Shi, X.Z., Du, K., Qiu, X.Y., Li, X.B., Mitri, H.S., “Feasibility of random- forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel,” Int. J. GeoMech., 17 (6), 04016129, 2017.

[14] Zhang, W.G., Zhang, Y.M., Goh, A.T.C., “Multivariate adaptive regression splines for inverse analysis of soil and wall properties in braced excavation,” Tunn. Undergr. Space Technol., 64, pp. 24-33, 2017.

[15] Goh, A.T.C., Zhang, W.G., Zhang, Y.M., Xiao, Y., Xiang, Y.Z., “Determination of earth pressure balance tunnel-related maximum surface settlement: a multivariate adaptive regression splines approach,” Bull. Eng. Geol. Environ. 77, 489-500, 2017.

[16] Xie, Q., Peng, K., “Space-time distribution laws of tunnel excavation damaged Zones (EDZs) in deep mines and EDZ prediction modeling by random forest regression,” Adv. Civ. Eng., pp. 1-13, 2019.

[17] Z. Chen, Y. Liu and S. Liu, “Mechanical state prediction based on LSTM neural network,” in Proc. of the 2017 36th Chinese Control Conference (CCC), Dalian, China, pp. 3876-3881, 2018.

[18] R. FukuoKa, H. Suzuki, T. Kitajima, A. Kuwahara and T. Yasuno, "Wind speed prediction model using LSTM and 1D-CNN," J-STAGE, 22(4), pp. 207-210, 2018.

[19] Y. H. Chen, T. Krishna, J. S. Emer and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE Journal of Solid-State Circuits, 52(1), 127-138, 2017

[20] F. J. Ordónez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," Sensors. pp. 3-9, 2016.

[21] T. D. Do, M. T. Duong, Q. V Dang and M. H. Le, "Real-time self-driving car navigation using deep neural network," in Proc. of the 2018 4th International Conference on Green Technology and Sustainable Development(GTSD), pp. 7-12, 2018.

[22] K. A. Althelaya, E. M. Alfy and S. Mohammed, "Evalutaion of bidirectional LSTM for short-and long-term stock market prediction," in Proc. of the 2018 9th International Conference on Information and Communication Systems(ICICS), Irbid, Jordan, pp. 151-156, 2018.

[23] D. Niu, Z. Xia, Y. Liu, T. Cai and Y. Zhan, "Alstm: Adaptiv LSTM for durative Sequential data," in Proc of the 2018 30th International Conference on Tools with Artificial Intelligence (ICTAI), Volos, Greece, pp. 151-157, 2018.

[24] A. Graves, N. Jaitly and A. R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in Proc. of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Repulic, pp. 273-278, 2013.

[25] Gers, F. A., N. N. Schraudolph, and J. Schmidhuber, Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research, pp. 115-143, 2002.

A Review on Classification Methods for Plants Leaves Recognition

Khaled Suwais[1]
Faculty of Computer Studies
Arab Open University (AOU)
Riyadh, Saudi Arabia

Khattab Alheeti[2], Duaa Al_Dosary[3]
College of Computer Sciences and Information Technology
University of Anbar
Anbar, Iraq

***Abstract*—Plants leaves recognition is an important scientific field that is concerned of recognizing leaves using image processing techniques. Several methods are presented using different algorithms to achieve the highest possible accuracy. This paper provides an analytical survey of various methods used in image processing for the recognition of plants through their leaves. These methods help in extracting useful information for botanists to utilize the medicinal properties of these leaves, or for any other agricultural and environmental purposes. We also provide insights and a complete review of different techniques used by researchers that consider different features and classifiers. These features and classifiers are studied in term of their capabilities in enhancing the accuracy ratios of the classification methods. Our analysis shows that both of the Support Victor Machines (SVM) and the Convolutional Neural Network (CNN) are positively dominant among other methods in term of accuracy.**

***Keywords*—Leaf recognition; feature extraction; leaf features; classifiers; image processing**

I. INTRODUCTION

An important role introduced by plants to maintain the ecological balance of the earth by providing us with breathing, shelter, fuel and medicine. Pattern recognition and image processing techniques are exploited by using plant images to build plant lists for the conservation and preservation of existing classes of the plant [1]. Leaves are considered convenient for the recognition and classification of different plant species because they are capable to present flat and two-dimensional surfaces with various characteristics like texture, colour, and shape. Many biological and environmental factors affect leaves to be damaged. So, many characteristics of a damaged leaf will be not useful to provide identifying signals. Therefore, a recognition system that depends on such characteristics may lead to unreliable and inconsistent outcomes.

Plant species recognition and classification method by conventional artificial processes are present time consuming, due to the depending on specific botanical information used by common persons [1]. Many research topics based on the automatic classification of the plant species are important. Some effective algorithms in computer science such as pattern recognition, image processing and machine learning and some technologies such, mobile devices and digital cameras, present the idea of automated classification for plant species by extracting different characteristics from the images of a plant leaf. With the development of machine learning, image processing, mobile devices, computer software, and hardware [2], it is possible to present an efficient and quick automated system to manage, recognize and understand a plant species [3].

In the area of plant taxonomy, leaf analysis has an essential role used to analyze, recognize and understand plant recognition and leaf patterns. The automatic plant recognition based on some features and characteristics, including leaf texture, leaf shape, leaf colour, and, other geometric features has been exploited. These characteristics are dependent on the recognition of the plant species. One of the essential challenges for plant recognition/classification is the diversity of leaf shapes [4]. The colour feature is more dependent to classify and identify plant species because leaf colour is can be changed according to the environment in different seasons. The texture features are more based on the information assured from its vein and venation. Recently, leaf venation patterns are considered an important factor to identify plant species with few techniques to extract leaf vein structure. Many methods depended on automatic or manual leaf venation extraction from leaf patterns. Furthermore, there have been few efforts to correlate and evaluate leaf venation and leaf spectral signatures [5].

In general, texture, shape, and colour features for each kind of plant leaf utilized to recognize plant species [4]. Therefore, most of the existing systems and methods of plant species recognition depend on these features of leaf image with its ability to be valid and reliable for years.

In this paper, different methods used in the plant recognition and classification field are discussed. The implementation and performance of various methods of plant recognition is important for the advancement of these technologies in supporting environment. Hence these methods are reviewed and analyzed. The presented methods have advantages and disadvantages for the recognition and identification of leaf patterns. The remainder of this paper organized as follows: Section 2 presents and discusses various earlier works. Section 3 presents the advanced methods used in leaf recognition. In Section 4, difficulties and directions related to the earlier proposed methods of leaf recognition are discussed. Conclusions are presented in Section 5.

II. Literature Review

In general, there is a general step for leaf recognition, including capturing leaf's images, applies pre-processing method on the captured image, extract feature and classify leaf. Fig. 1 illustrates the flowchart of the major steps carried out in the process of leaf recognition.

A. Images Capturing

In various studies, a scanner or digital camera is used for acquiring leaf images. In [6], the authors used a Samsung camera (DV300F SAMSUNG zoom Lens 5X 16.1 megapixels) to capture images of on-branch green apples, apricot, nectarine, sour cherry, peach, and amber-coloured plums. A digital camera (SONY W730) is used in [7] to capture the green apple targets. The Microsoft Kinect 2.0 camera is chosen in [8] to capture juicy peach images for colour, depth, and point cloud features. While the authors in [9] used an MX808 camera to collect green pepper plant images to create a new dataset. The Canon 660D digital camera used to collect 8911 images of rice leaf disease as a dataset used in the paper [10].

To collect 2D images for apple fruit counting and diameter, the authors in [11] used a thermal camera for accurate results. Also, a thermal camera is used in [12] to collect 2D images of oranges for recognition. Because of the limitation presented with 2D images related to incomplete information, 3D images are considered in many types of research. A laser scanner used in [13], [14], [15] to scan 3D images. Alternatively, an RGB-D camera is used in [16], [17], [18] to present a complete and significantly 3D scan.

B. Images Pre-Processing Methods

An important concept in the leaf recognition system is the pre-processing phase. This phase includes the following steps: image re-orientation, image cropping, convert the image to a grayscale image than to a binary image, remove noise, stretch contrast, and threshold inversion [19]. Various preprocessing techniques are developed based on efficient machine learning methods. How leaf images' features are extracted, and the outcomes of pre-processing phase are important aspect of visual-based machine learning. The study in [20], suggested that to extract leaf features, the leaf image is divided into 2/4 parts, instead of the whole leaf extraction. Vein, colour, Fourier descriptors (FD) exploited in the presented image processing techniques. To achieve a sufficient rate of accuracy, Gray-Level Co-occurrence Matrix (GLCM) methods and the Flavia leaf dataset are used to present 99.1% accuracy. In [21], presented a study and analysis of different methods used various image pre-processing techniques. Simple Linear Iterative Clustering (SLIC) used in one of the studied methods, which uses on super-pixel for grouping them with a defined value through many iterations of the closed neighbour to determine a data vector with a similar value. In [22], the Guided Active Contour (GAC) method is developed. In this method, the snake segmentation technique is used to enhance the polygonal framework for the elongated leaf shape.

For extraction of segments from the data, a hierarchical model based on the Kurtz algorithm is proposed [23]. The proposed approach suggests extracting the interesting parts from data. The data is arranged from the lowest to the highest resolution as clusters as a tree. The first cluster represents the colour features of coarse image patches. The Binary Partition Tree (BPT) used to arrange the individual patches in a hierarchical manner. This method shows that the precision of the system reached up to 85.1%. In [24], a pre-processing technique is used in the proposed system for recognition of soybean and weed leaf. The data used include the images captured by the 2G-R-B camera where the erosion algorithm utilized to remove images distortion. Moment invariant is used to identify scale, invariability, rotation, and translation of soybean leaf image. The image pre-processing technique used can improve the classification rate to 90.5%.

C. Feature Extraction Methods

Some important characteristics such as colour, size, and shape are used for leaf recognition. The segmented image can be a source of information for feature extraction and could assist in the proper classification of the anomaly. Some statistical measures used for textural features extraction such as Color Co-occurrence Matrix (CCM), Spatial Grey Level Dependence Matrix (SGLDM), Grey Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP). Various existing systems and methods of plant recognition depend on the colour, size, shape, and texture of the leaf image.

The study in [25] leaves in plants have holes or diseases that could cause reduction of leaves, and thus cause segmentation. First, point searched by pixel scanning and arranged as foreground/background. When the pixel is categorized as foreground, this process cuts off and the next line is scan. Every individual pixel passed with this process for identification. The result of this model was provided with an average error of 3.00 for five leaves. The study in [26] proposed a system for applying feature extraction by utilizing a method known as area labelling. The pre-processing phase is applied for image processing to provide binary image output. Next, the output binary image is offered to area labelling for identified region production. In this work, when the pointer defines a pixel with the value '1' then the eight-connecting area algorithm is used to acquire more search for the eight-connecting area by the kernel. The features of the leaf image are reflected when the pixels are marked and contend for features extraction.

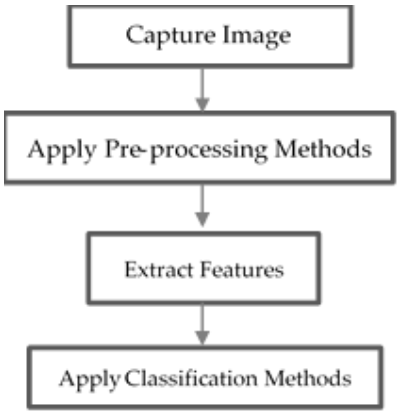


Fig. 1. The Basic Steps of Leaf Recognition Method.

In [27], Gopal et al. attended to present the medicinal images for classification goals depending on the colour features extraction. First, a digital scanner used for providing input image attached for the preprocessing phase. Later, for feature extraction, the image is pushed into the program to gain the colour feature according to its Fourier descriptor. In the training phase, 100 leaf images are used and 50 leaf images are used in the testing phase. The results show that the efficiency of the method is 92%. Feature extraction represents an important role in providing accurate precision and accuracy in leaf recognition/classification system built on utilizing machine learning mechanism. This belongs to the fact that the predetermined feature in the network affects the architecture of machine learning. Different mechanisms used in different approaches to solving different problems so that there are various feature extraction methods to be utilized.

III. Leaf Recognition and Classification Methods

Various related researches proposed for leaf plant recognition and classification is discussed in this section. In [28], the Local Binary Patterns (LBP) method is used to propose an alternative method for plant leaves classification. The proposed method uses the extracted texture features from plant leaves to recognize plant leaves. LBP, the R and G colour of images. In addition, the method efficiency against Gaussian, pepper, and salt are evaluated. Next, the Extreme Learning Machine (ELM) method is used to classify and test the acquired features from the proposed system. In this system, Swedish, Flavia, Foliage, and ICL datasets are used. The obtained results are compared to prove that the proposed method can identify noiseless from noisy images. The accuracy results achieved is claimed to be (98.94%) Flavia, (99.46%) Swedish, (83.71%) ICL and (92.92%) Foliage datasets.

An automatic and accurate segmentation method is proposed in [29]. The authors have used an efficient encoding method for the feature depth information extraction. Later, Mask R-CNN is deployed to train the used RGB-D data. For more efficiency, the features of the data are fused in the Feature Pyramid Network (FPN) structure. Next, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) provided to segment a single leaf from overlapping leaves in the explored scope using the detected leaf areas and depth data. The experimental results are compared to prove that the proposed system automatically detects leaves with an accuracy of around 89.3%. In [30], the authors used the dataset of apple leaf image that employed six apple leaf diseases to provide 2462 images for method evaluation. The proposed method is compared with the traditional multi-classification method based on cross-entropy loss function for results evaluation. The traditional multi-classification method achieves an accuracy of 92.29%, while the proposed method in [30] presents better accuracy with 93.51%, 93.31%, and 93.71% on the test set, respectively.

In [31], Jaya Algorithm with the optimized deep neural network used to propose a system for paddy leaf diseases identification. The leaves image of the rice plant is taken normally from the field, brown spot, blast, and sheath rot diseases. In the pre-processing phase, the RGB images are converted into HSV images and binary images are extracted to split the non-diseased and diseased samples. For the segmentation of non-diseased portion, diseased portion, and background a clustering method is utilized. Jaya Optimization Algorithm (DNN_JOA) with Optimized Deep Neural Network is used in the Classification of diseases phase. The results of the work prove that the proposed method achieved an accuracy of 90.6%.

The authors in [32] presented a classification method of plant's leaves based on Multiscale Triangle Descriptor (MTD) and Local Pattern Histogram Fourier (LBP-HF). The two methods are employed to characterize shape and texture, respectively. Based on their experiments, the recognition accuracy ratio is found to be 99.1%, 98.4%, 95.6% when applied on Flavia, Swedish and MEW2012 datasets, respectively. However, the method has some limitations. The features of the leaves need to be designed manually as no automated process of learning is applied. In [33], an alternative recognition method is presented based on Generalized Procrustes Analysis (GPA). The method uses contour (shape) features for classification. The core of the method depends on performing some computation to calculate the distance between a set of contour points and the center of the contour upon applying some alignments. The results show that the recognition accuracy rate is 84.4% and 98.4% on Leafsnap and Flavia datasets, respectively.

A recognition method based on Multiscale Sliding Chord Matching (MSCM) is presented in [34]. The method aims to recognize soybean cultivar by joint leaf patterns. The MSCM strategy is implemented to extract shape features. The experiment over 6000 sample images shows that the accuracy ratio is 72.4%. The analysis shows that such a low ratio results from several reasons. The leaves of the soybean plan have different visual cues for soybean cultivar identification. In addition, the joint leaf pattern is not integrated with the descriptors of leaves from different parts of soybean plants. There are many other classification models found in various researches. These methods include Support Vector Machine, Artificial Neural Network, Convolutional Neural Network, K-Nearest Neighbors, and Probabilistic Neural Network.

A. Support Vector Machine (SVM)

In SVM is an essential machine-learning technique for data learning and solving classification and identification problems. The study in [35] proposed utilizing leaf contour and centroid for proposing the leaf image recognition systems. The proposed method aimed to use image processing techniques as well as SVM utilized as a classifier. Flavia dataset utilized to take 70 patterns with their shape and geometrical features. Their results prove that the highest achievement accuracy of 97.7%. In [36], the authors provide a comparative analysis for leaf recognition and classification. SVM used as the classifiers in this system and a shape detector utilized to extract 14 leaf features. In the training dataset, the Flavia database used to provide sixteen different plant species. The results show that the highest accuracy of 90.9% by exploiting SVM.

Araujo et al. [37] used SVM and neural network as classifiers of leaf image classification. These classifiers used

for training four different features, the histogram of gradients (HOG), namely local binary pattern (LBP), Zernike Moments (ZM), and speed of robust features (SURF). The results show that using multiple classifiers of the system overcame the performance of monolithic methodologies and the best results reported. A significant improvement proved to be effective to detect plants by using SVM as a classifier for an environment with heavy overlapping and interferences cases [38]. In this experiment, the authors exploited 300 leaf images of three plant species for identification. A marker-controlled watershed segmentation was used to capture and segment the images. The system achieves 86.7% accuracy for identification. The accuracy can be improved by adding more features as well as the dataset used for the experiments. SVM suffer from different limitations, such as the complexity of its structure, and the slowness of training and testing. On the other hand, SVM is considered robust and has high potentials for generalization.

B. Artificial Neural Network (ANN)

The proposed system of leaf pattern recognition in [39] exhibits that using ANN as a classifier is reliable. There was a study presents 98.6% of accuracy for recognition which can be increased when more dataset used [39]. In [40], using ANN as a classifier to recognize and identify the medicinal plant leaves can improve the results. The ANN classifier was used to train the extracted colour, shape, and texture of leaf images. The results show that the system presents an accuracy of 94.4% using 63 leaf images. The accuracy of the extracted leaf venation improved in [41] by about 10% when selecting the ANN as classifier combined with thresholding. The results show that the accuracy improved to 97.3% by combining ANN with thresholding. ANN can recognize the relationships between dependent/independent variable, and support simplistic statistical testing. As for the limitations, ANN requires a high computational load and a high tendency of data overfitting.

C. Convolutional Neural Network (CNN)

In [42], CNN is used to establish a cotton growth recognition algorithm. Confusion matrix and recognition efficiency exploited for the optimization process where a CNN model is established, and its precision was proved by modifying training /test sets based on the concept of the k-fold test. The results show that this method is suitable for the recognition task and can achieve good results in the term of high precision, low cost, and real-time. The method proposed in [43] presents an automated system for medicinal plant classification using CNN. A 3-layer CNN is employed to extract high-level features for classification. The method is supported by a data augmentation technique for higher efficiency. The experimental results show that the recognition accuracy rate of the method is around 71.3%.

To solve the disease similarity problem, an efficient method is proposed in [44]. Two types of diseases happening in the same leaf and the influence of external light lead to this problem. In the beginning, they gained a cucumber leaf disease dataset, then they build a classification model by using the EfficientNet method for the above four types. Finally, they used CNN-based EfficientNet-B4 to demonstrate a two-classification model of cucumber similar diseases. The obtained results prove that their proposed method has a considerable effect on the similar diseases of cucumber classification of accuracy around 96%. In [10], the authors used CNNs to extract the rice leaf disease image features. Later, for classification and prediction of the specific disease SVM method is applied. In their work, the cross-validation method was the optimal parameter of SVM. The results show that the average accuracy of the proposed recognition model was 96.8% based on utilizing deep learning and SVM techniques. The experiment is applied over a dataset prepared by the authors as per the details stated in Table I.

In [45], a deep convolutional neural network used to build an automatic classification and recognition framework of various paddy crop stress as biotic/ abiotic using the field images. The dataset used includes 12 different stress categories of healthy/normal with 30,000 field images of five different paddy crop varieties. The results show that the proposed model can achieve an average accuracy of 92.89%. In image recognition tasks, CNNs are used as feature extractors and classifiers to introduce the better performance. In CNN's, Multiple features are extracted simultaneously as well as they are robust to noise. These advantages made CNN an interesting classifier in many types of research. In [46], the authors aimed to identify leaf diseases based on the traditional CNN by integrating of inception structure and a pooling layer. In this model, the number of parameters reduced and the identification accuracy improved by up to 91.7%. Similarly, the model in [47] used CNN classifier for maize leaf disease detection. This method can classify diseases according to three types. For plant disease classification and recognition, CNN is proven to be an effective manner. The method in [48] integrates deep learning with CNN for classification. The results show that even reducing the number of parameters, would not affect the recognition accuracy.

CNN considered a faster recognition process as it extracts and recognizes the features concurrently. CNN is accurate for plant classification due to the numerous sets of data trained by users before it is considered to be capable enough for application. CNN shows that the accuracy of leaf classification achieved up to 94% [49]. The integration of deep learning knowledge with CNN provided an efficient model for feature extraction to recognize and identify vein samples from the presented image [50].

D. K-Nearest Neighbors (KNNs)

In the recognition and classification methods, the accuracy of identification increased when the number of images for testing is increased. The study in [51] shows that Principal Component Analysis (PCA) algorithm and Cosine k-Nearest Neighbors (KNN) classifier is improved compared to SVM and Patternnet neural network. KNN classifier provides 83.5% of accuracy [19]. Such low accuracy is relatively weak to be agreeable even the process of feature extraction is quick and simple. KNN classifier is not capable to handle samples distortion and could cause inaccuracy in the classification process. A method proposed for this classifier with a specific colour histogram increases the accuracy up to 87.3% [19].

TABLE I. PREVIOUS LEAF CLASSIFICATION METHODS

Ref.	Published year	Dataset	Classifier	Extracted Features	Average Accuracy Rates
[28]	2019	Flavia *dataset* Swedish *dataset* ICL *dataset* Foliage *dataset*	LBP	Color Texture	Flavia = 98.94% Swedish = 99.46% ICL = 83.71% Foliage = 92.92%
[29]	2020	7988 *images*	Mask R-CNN	Color	89.03%
[31]	2019	650 *images*	Deep Neural Network	Color Texture	90.57%
[32]	2021	Flavia *dataset* Swedish *dataset* MEW2012 *dataset*	MTD + LBP-HF	Texture Shape	Flavia = 99.10% Swedish = 98.40% MEW2012 = 95.60%
[33]	2018	Leafsnap *dataset* Flavia *dataset*	GPA	Shape	Leafsnap = 84.40% Flavia = 98.40%
[34]	2020	6000 *images*	MSCM	Shape	72.40%
[35]	2017	Flavia *dataset*	SVM	Shape	97.70%
[36]	2018	Flavia *dataset*	SVM	Shape	90.90%
[38]	2015	300 *images*	SVM	Shape	86.70%
[37]	2017	ImageCLEF 2011 *dataset* ImageCLEF 2012 *dataset*	SVM + Neural Network	Texture Shape	ImageCLEF 2011 = 86.20% ImageCLEF 2012 = 64.10%
[10]	2020	8911 *images*	SVM + CNN	Shape Color	96.80%
[45]	2020	6000 *images*	CNN	Shape	92.89%
[30]	2020	2462 *images*	CNN	Color Texture Shape	92.29%
[44]	2020	2816 *images*	CNN	Color Texture Shape	96.00%
[42]	2020	1443 *images*	CNN	Texture	93.27%
[46]	2019	6108 *images*	CNN	Color Texture	91.70%
[47]	2019	54306 *images*	CNN	Texture	92.85%
[48]	2019	ImageNet *dataset* PlantVillage *dataset*	CNN	Color	97.14%
[49]	2017	Flavia *dataset*	CNN	Shape	99.70%
[43]	2020	3570 *images*	CNN	Shape Vein	71.30%
[39]	2006	180 *images*	ANN	Shape Vein	94.40%
[40]	2013	63 *images*	ANN	Shape Color Texture	94.40%
[41]	2007	2940 *images*	ANN	Color Vein	97.33%
[51]	2019	ImageCLEF 2012 *dataset* Leafsnap *dataset* Flavia *dataset*	KNNs	Texture	ImageCLEF 2012 = 88.80% Leafsnap = 74.50% Flavia = 98.70%
[52]	2016	Flavia *dataset*	KNNs	Shape	94.37%
[53]	2010	1200 *images*	PNN	Shape	91.41%
[54]	2008	900 *images*	PNN	Shape Texture	93.70%
[55]	2014	Flavia *dataset* Swedish *dataset*	PNN	Shape	Flavia = 82.01% Swedish = 80.01%
[56]	2012	2448 *images*	PNN	Texture Color	74.51%.
[57]	2007	1800 *images*	PNN	Texture Color Shape	90.00%

The authors in [52] produced an improvement in leaf classification based on utilizing KNN classifier with edge and shape features. Flavia dataset exploited to provide 32 plant species to be tested. The results show that the presented method improves the average classification accuracy to 94.4%.

E. Probabilistic Neural Network (PNN)

In the recognition and classification methods, PNN is utilized as a classifier due to many advantages, including high resistance of distortion, flexibility to modify data, and the specimen can be classified into multiple outputs. In this section, we study the efficiency of PNN in classifying leaves.

The work in [53] presents an algorithm for plant species classification of leaf image based on PNN. The points of the leaf's shape are extracted from the background and a binary image is produced accordingly. After that, the leaf is aligned horizontally with its base point on the left of the image. Several morphological features, such as eccentricity, area, perimeter, major axis, minor axis, equivalent diameter, convex area and extent, are extracted. The network was trained with 1200 simple leaves from 30 different plant species with an accuracy rate of 91.41%. The authors in [54] address the issue of low recognition rate in plant identification since the objects broad and the classification features are not synthetic. To resolve this issue, PNN is presented for a rapid recognition method that is applied over thirty kinds of broad-leaved trees. The shape and texture features of broad-leaved trees combine, composing a synthetic feature vector of broad leaves to realize the computer automatic classification towards broad-leaved plants. The use of PNN has achieved an average recognition rate of 93.70%.

An alternative PNN-based leaf classification method is proposed in [55]. Upon converting the RGB image to its binary image representation, the binary image is passed to a canny operator to recognize the edges of the image. Sampling is then used to compute the centroid distance of these points and the distance of sampling points from the axis of the least inertia line. A probabilistic neural network has been used as a classifier. The results show that the average accuracy rates of the method on Flavia and Swedish datasets are 82.1% and 80.1%, respectively. In [56], the researchers present a mobile application for identifying Indonesian medicinal plants. The application uses both Fuzzy Local Binary Pattern (FLBP) and Fuzzy Color Histogram (FCH) methods for extracting leaf image texture and colour, respectively. For fusion of FLBP and FCH, the Product Decision Rules (PDR) method is applied. As for the classifier, PNN is utilized to classify medicinal plant species. The accuracy of this work is claimed to be around 74.51%. PNN appear to be an effective classifier for the automated leaf recognition method proposed in [57]. The method relies on the use of image and data processing techniques, and applied over 1800 leaf images. The method managed to extract 12 leaf features organized into 5 basic variables which compromise the PNN input vector. The PNN is trained by 1800 leaves to classify 32 kinds of plants. The accuracy is found to be reasonable around 90%. However, aside from the advantages of PNN mentioned above, PNN is considered as a complicated network layout, and it requires long time on training. In addition, PNN has a tendency for overfitting with too many traits. Table I summarizes the key facts and finding of our analysis.

IV. Discussion and Analysis

In the early presented plants leaves species recognition systems, several issues related to providing better classification results are addressed. Our analysis of existing classification methods focuses on different issues, including the commonly used features and classifiers and their impact on classification accuracy, what datasets are used for testing, and research trends on leaf classification methods.

Researchers have used several features in their methods, including (colour (C), shape (S), texture (T) and vein (V)). We have also found some researches combine multiple features to enhance the accuracy ratio. Most of the researches ($\approx$ 41% of existing methods) focus on shapes features in their classification methods. Analysis of the accuracy ratio of these methods shows that combining multiple features in the classification method helps in enhancing the accuracy rations of leaves classifications. Our analysis also reveals that there is a lack of studies on methods that use vein as a feature of classification, as only $\approx$ 6% of existing studies tickle such feature in their methods. However, considering the vein features shows promising results when combined with shape, colour or texture features. Fig. 2(a) shows the percentage of studies discuss each type of features, while Fig. 2(b) reflects the accuracy ratios achieved by these features according to the existing classification methods.

As for classifiers, various techniques are found in the state of the art. We found that there is a greater focus on CNN-based classifiers. Several methods show enhanced performance when combining CNN with other classifiers, such as SVM and LBP. Most of the accuracy ratio shows that CNN-based methods outperform other classifiers. On the other hand, there is a growing interest in ANN classifiers as it shows high accuracy ratios. In the three existing studies on ANN classifiers, results show that the accuracy ratio ranges between 94.4 and 97.3. Such high accuracy should give ANN classifier more interest for researchers in developing new classification methods. Fig. 3 presents the accuracy ratio achieved by different classifiers.

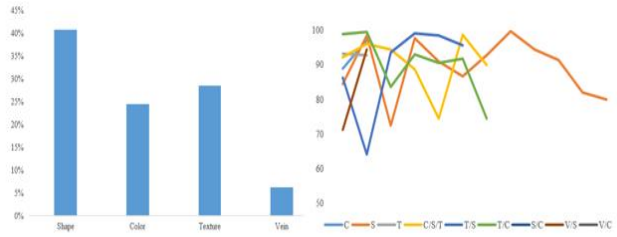


Fig. 2. (Left) Appearance Ratio of Features in Existing Methods, (Right) Accuracy Ratio of different Features.

In term of testing environments, our analysis shows that the majority of researchers (≈ 48%) has developed their datasets for their testing. Using well-known standard datasets such as Flavia and Swedish appeared in less than 30% of the studies. In this regard, researchers should focus on updating and considering standard datasets to enhance the scientific judgments on proposed classification methods. Fig. 4 shows the utilization of different datasets for testing leaves classification methods.

As for the classification method, we noticed that the current researches are oriented toward three main areas. These areas are CNN, SVM and PNN. We found that CNN occupies ≈ 31% of existing classification methods, while each of PNN and SVM found in ≈ 16% of methods. Fig. 5 illustrates the frequencies of different classification methods used in the state of arts.

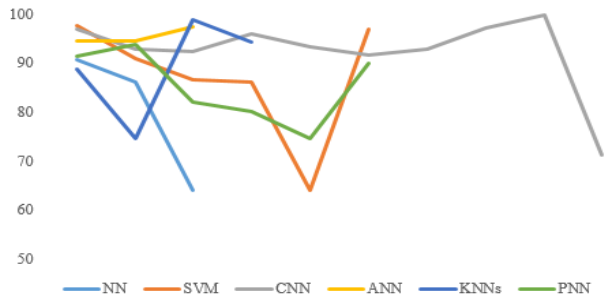


Fig. 3. Accuracy Ratios Achieved by different Classifiers.

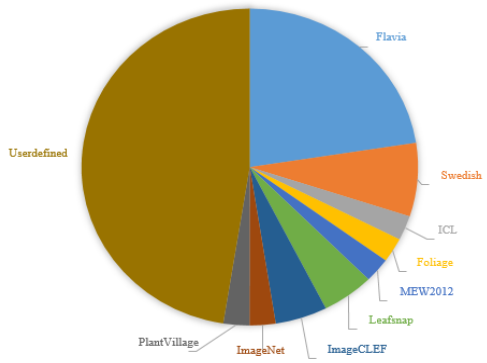


Fig. 4. Datasets used for Testing Leaves Classification Methods.

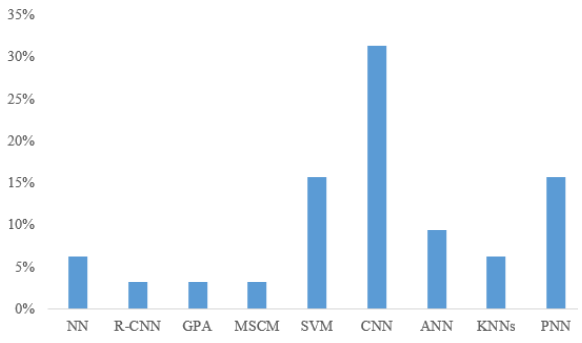


Fig. 5. Classification Methods used for Leaves Classification and Recognition.

V. Conclusion

In this research, we have made an effort to study and analyze the latest researches in the field of leaves classification and recognition. We have provided helpful insight on the process of leaves classification using different features of leaves. These features are discussed and analyzed thoroughly, and their efficiency in enhancing the recognition and classification process is presented. In addition, various classifiers and classification methods are studies. Our unique analysis has discussed and analyzed different factors that might affect the accuracy of the classification process. These factors include features, classifiers, and testing datasets. We found that combining multiple features have a positive impact on the classification process. However, greater efforts should be made by researchers to examine and investigate the best combination of features. For instance, none of the researches has combined the vein feature with the colour, shape and texture in one method. We found that CNN classifiers groups the attention of researchers, while SVM classifiers are found more attractive in recent researches. As SVM classifiers look interesting in recent years, further investigations are needed to study the relation between accuracy and the best leaves' features that should be used in SVM-based classification methods. As for the testing datasets, we found that more efforts should be made on unifying these datasets for integrity purposes. The majority of researchers have tested their methods based on some user-defined datasets, which makes the comparisons between proposed methods inaccurate.

Acknowledgment

The authors would like to thank the Arab Open University, Saudi Arabia for supporting this study.

References

[1] Y. Shao, "Supervised global-locality preserving projection for plant leaf recognition," Computers and electronics in agriculture, vol. 158, no. October 2018, pp. 102–108, 2019.

[2] J. Zhang, D. Huang and T. Lok, "A novel adaptive sequential niche technique for multimodal function optimization," Neurocomputing, vol. 69, pp. 2396–2401, 2006.

[3] M. Seeland, M. Rzanny and N. Alaqraa, "Plant species classification using flower images—A comparative study of local feature representations," PLoS one, vol. 12, no. 2, 2017.

[4] Q. Zhao, H. Ma. and M. Cheung, "An efficient android-based plant leaf identification system," Neurocomputing, vol. 151, pp. 1112–1119, 2015.

[5] S. Green, A.Walton, S. Little, C. Price, S. Wing et al., "Reading the leaves: a comparison of leaf rank and automated areole measurement for quantifying aspects of leaf venation," Applications in plant sciences, 2, 2014.

[6] S. Iman, and H. Khosravi, "Expert systems with applications a deep neural network approach towards real-time on-branch fruit recognition for precision horticulture," Expert systems with applications, vol. 159, p. 113594, 2020.

[7] S. Sun, Q. Wu, L. Jiao, Y. Long, D. He et al., "Recognition of green apples based on fuzzy set theory and manifold ranking algorithm," International journal for light and electron optics, 2018.

[8] G. Wu, B. Li, Q. Zhu, M. Huang and Y. Guo, "Using color and 3d geometry features to segment fruit point cloud and improve fruit recognition accuracy," Computers and electronics in agriculture, vol. 174, no. January, p. 105475, 2020.

[9] W. Ji, X. Gao, B. Xu, G. Chen and D. Zhao, "Target recognition method of green pepper harvesting robot based on manifold ranking," Computers and electronics in agriculture, vol. 177, no. July, 2020.

[10] F. Jiang, Y. Lu, Y. Chen, D. Cai and G. Li, “Image recognition of four rice leaf diseases based on deep learning and support vector machine,” Computers and electronics in agriculture, vol. 179, no. October, p. 105824, 2020.

[11] D. Stajnko, M. Lakota and M. Hočevar, “Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging,” Computers and electronics in agriculture, vol. 1, pp. 31–42, 2004.

[12] D. Bulanon, T. Burks and V. Alchanatis, “Image fusion of visible and thermal images for fruit detection,” Biosystems engineering, vol. 113, pp. 12–22, 2009.

[13] P. Eizentals and K. Oka, “3D pose estimation of green pepper fruit for automated harvesting,” Computers and electronics in agriculture, vol. 128, pp. 127–140, 2016.

[14] J. Mack, C. Lenz, J. Teutrine and V. Steinhage, “High-precision 3D detection and reconstruction of grapes from laser range data for efficient phenotyping based on supervised learning,” Computers and electronics in agriculture, vol. 135, pp. 300–311, 2017.

[15] S. Paulus, J. Dupuis, A. Mahlein and H. Kuhlmann, “Surface feature based classification of plant organs from 3D laserscanned point clouds for plant phenotyping,” BMC bioinformatics, vol. 14, p. 238, 2013.

[16] E. Barnea, R. Mairon and O. Ben-Shahar, “Colour-agnostic shape-based 3D fruit detection for harvesting robots,” Biosystems engineering, vol. 146, pp. 57–70, 2016.

[17] R. Perez, F. Cheein and J. Rosell-Polo, “Flexible system of multiple RGB-D sensors for measuring and classifying fruits in agri-food Industry,” Computers and electronics in agriculture, vol. 139, pp. 231–242, 2017.

[18] T. Yongting and Z. Jun, “Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking,” Computers and electronics in agriculture, vol. 142, pp. 388–396, 2017.

[19] T. Munisami, M. Ramsurn, S. Kishnah and S. Pudaruth, “Plant leaf recognition using shape features and colour histogram with k-nearest neighbour classifiers,” Procedia computer science, vol. 58, pp. 740 – 747, 2015.

[20] M. Turkoglu and D. Hanbay, “Recognition of plant leaves: An approach with hybrid features produced by dividing leaf images into two and four parts,” Applied mathematics and computation, vol. 352, pp. 1–14, 2019.

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua et al., “SLIC superpixels compared to state-of-the-art superpixel methods,” IEEE transactions on pattern analysis and machine intelligence, vol. 34, pp. 2274–2281, 2012.

[22] G. Cerutti, L. Tougne, J. Mille, A. Vacavant and D. Coquin, “Understanding leaves in natural images—a model-based approach for tree species identification,” Computer vision and image understanding, vol. 117, pp. 1482–1501, 2013.

[23] C. Kurtz, N. Passat, P. Gançarski and A. Puissant, “Extraction of complex patterns from multiresolution remote sensing images: a hierarchical top-down methodology,” Pattern recognition, vol. 45, pp. 685–706, 2012.

[24] Z. Bo, W. Hua, L. Jun, M. Hua and Z. Chao, “Research on weed recognition method based on invariant moments,” 11 world congress on intelligent control and automation, pp. 2167–2169, 2014.

[25] C. Lu, H. Ren, Y. Zhang and Y. Shen, “Leaf area measurement based on image processing,” 9th international conference on measuring technology and mechatronics automation, vol. 2, pp. 580–582, 2010.

[26] V. Shivling, A. Singla, C. Ghanshyam, P. Kapur and S. Gupta, “Plant leaf imaging technique for agronomy,” 2011 International conference on image information processing, pp. 1–5, 2011.

[27] A. Gobal, S. Reddy and V. Gayatri, “Classification of selected medicinal plants leaf using image processing,” 2012 International conference on machine vision and image processing, pp. 5–8, 2012.

[28] M. Turkoglu and D. Hanbay, “Leaf-based plant species recognition based on improved local binary pattern and extreme learning machine,” Physical A: statistical mechanics and its applications, vol. 527, p. 121297, 2019.

[29] X. Liu, C. Hu and P. Li, “Automatic segmentation of overlapped poplar seedling leaves combining mask R-CNN and DBSCAN,” Computers and electronics in agriculture, vol. 178, no. August, p. 105753, 2020.

[30] Y. Zhong and M. Zhao, “Research on deep learning in apple leaf disease recognition,” Computers and electronics in agriculture, vol. 168, no. October 2019, p. 105146, 2020.

[31] S. Ramesh and D. Vydeki, “Recognition and classification of paddy leaf diseases using optimized deep neural network with Jaya algorithm,” Information processing in agriculture, vol. 7, no. 2, pp. 249–260, 2020.

[32] C. Yang, "Plant leaf recognition by integrating shape and texture features", Pattern recognition, vol.112, p. 107809, 2021.

[33] S. Choudhury, J. Yu and A. Samal, "Leaf recognition using contour unwrapping and apex alignment with tuned random subspace method", Biosystems engineering, vol. 170, pp. 72-84, 2018.

[34] B. Wang, Y. Gao, X. Yuan and S. Xiong, "From species to cultivar: Soybean cultivar recognition using joint leaf image patterns by multiscale sliding chord matching", Biosystems engineering, vol. 194, pp. 99-111, 2020.

[35] A. Khmag, S. Al-Haddad and N. Kamarudin, “Recognition system for leaf images based on its leaf contour and centroid,” 2017 IEEE 15th student conference on research and development, pp. 467–472, 2017.

[36] V. Srivastava and A. Khunteta, “Comparative analysis of leaf classification and recognition by different SVM classifiers,” 2018 International conference on inventive research in computing applications, pp. 626–631, 2018.

[37] V. Araujo, A. Britto, A. Brun, A. Koerichm and R. Palate, “Multiple classifier system for plant leaf recognition,” 2017 IEEE international conference on systems, man and cybernetics, pp. 1880–1885.

[38] R. Nesaratnam and C. Murugan, “Identifying leaf in a natural image using morphological characters,” International conference on innovations in information, embedded and communication systems, 2015.

[39] Q. Wu, C. Zhou and C. Wang, “Feature extraction and automatic recognition of plant leaf using artificial neural network,” Research on computing science, vol. 20, pp. 3–10, 2007.

[40] R. Janani and A. Gopal, “Identification of selected medicinal plant leaves using image features and ANN,” 2013 International conference on advanced electronic systems, pp. 238–242, 2013.

[41] H. Fu and Z. Chi, “Combined thresholding and neural network approach for vein pattern extraction from leaf images,” IEE Proc. - Vision, image signal process, vol. 153, pp. 881–892, 2007.

[42] S. Wang, Y. Li, J. Yuan, L. Song and X. Liu, “Recognition of cotton growth period for precise spraying based on convolution neural network,” Information processing in agriculture, pp. 1–13, 2020.

[43] R. Akter and M. Hosen, "CNN-based leaf image classification for Bangladeshi medicinal plant recognition", Emerging technology in computing, communication and electronics, pp. 1-6, doi: 10.1109/ETCCE51779.2020.9350900.

[44] P. Zhang, L. Yang and D. Li, “EfficientNet-B4-ranger: a novel method for greenhouse cucumber disease recognition under natural complex environment,” Computers and electronics in agriculture, vol. 176, no. July, p. 105652, 2020.

[45] B. Anami, N. Malvade and S. Palaiah, “Deep learning approach for recognition and classification of yield affecting paddy crop stresses using field images,” Artificial intelligence in agriculture, vol. 4, pp. 12–20, 2020.

[46] B. Hang, D. Zhang, P. Chen and J. Zhang, “Classification of plant leaf diseases based on improved convolutional neural network,” Sensors, vol. 19, p. 4161, 2019.

[47] M. Sibiya and M. Sumbwanyambe, “A Computational procedure for the recognition and classification of maize leaf diseases out of healthy leaves using convolutional neural networks,” AgriEngineering, vol. 1, pp. 119–131, 2019.

[48] Y. Toda and F. Okura, “How convolutional neural networks diagnose plant disease,” Plant phenomics, vol. 2019, Article ID 9237136, doi: 10.34133/2019/9237136.

[49] W. Jeon and S. Rhee, "Plant leaf recognition using a convolution neural network," Korean institute of intelligent systems, vol. 17, pp. 26–34, 2017.

[50] G. Guillermo, L. Uzal, M. Larese and P. Granitto, "Deep learning for plant identification using vein morphological patterns," Computers and electronics in agriculture, vol. 127, pp. 418–424, 2016.

[51] F. Kheirkhah and H. Asghari, "Plant leaf classification using GIST texture features," IET computer vision, vol. 13, p. 369, 2018.

[52] P. Kumar, K. Rao, A. Raju, and D. Kumar, "Leaf classification based on shape and edge feature with k-NN classifier," 2016 2nd Int. Conf. Contemp. Comput. Informatics, pp. 548–552, 2016.

[53] J. Hossain, and M. Amin, "Leaf shape identification based plant biometrics," Proc. 2010 13th Int. Conf. Comput. Inf. Technol. ICCIT 2010, no. Iccit, pp. 458–463, 2010.

[54] L. Huang, and P. He, "Machine recognition for broad-leaved trees based on synthetic features of leaves Using Probabilistic Neural Network," Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008, vol. 4, pp. 871–877, 2008.

[55] K. Mahdikhanlou, and H. Ebrahimnezhad, "Plant leaf classification using centroid distance and axis of least inertia method," 2014 22nd Iran. Conf. Electr. Eng., pp. 1690–1694, 2014.

[56] Y. Herdiyeni, and N. Wahyuni, "Mobile Application for Indonesian Medicinal Plants Identification using Fuzzy Local Binary Pattern and Fuzzy Color Histogram," 2012 Int. Conf. Adv. Comput. Sci. Inf. Syst., pp. 978–979, 2012.

[57] S. Wu, F. Bao, E. Xu, Y. Wang, Y. Chang, and Q. Xiang, "A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network," IEEE Int. Symp. Signal Process. Inf. Technol., 2007.

Tracking Axonal Transports in Time-Lapse Images Obtained from a Microfluidic Culture Platform

Nak Hyun Kim
Division of Computer Engineering
Hankuk University of Foreign Studies, Kyunggi-do, South Korea

***Abstract*—In this paper, a procedure is described for tracking moving object trajectories from image sequences acquired from a microfluidic culture platform. Since particles move along the axons, curve structures need to be detected first from the input image sequence. A kymograph analysis technique is applied to detect axon structures from the consolidated image of the input sequence. Horizontally and vertically oriented axons are then detected by applying the process twice to the original and the 90-degree rotated image. Multiple kymographs are generated along the detected axons by projecting image intensity variation through the time-axis. The trajectory detection process is then applied to each kymograph image. To obtain the particle motion information from the entire image sequence, an integration process is applied to each horizontal and vertical kymograph data set. The proposed technique has been applied to image sequences in the present application area. It is demonstrated that practical results can be obtained using time-lapse image sequence data.**

Keywords*—Axonal transports; kymograph; trajectory detection; image sequence analysis; motion parameter extraction

I. INTRODUCTION

Time-lapse video of cellular motion in microfluidic culture platform consists of microscopic images acquired at fixed time intervals [1],[2],[3]. In fluorescence microscopic images, it is important to trace cellular motions appearing as a collection of irregular movements of small particles. While most organelles occupy stationary positions through time, some particles exhibit motion. The purpose of the analysis of mitochondrial transports is to identify motion trajectories and to detect the amounts of motions and speeds of particles. Tracking mitochondria in neural time-lapse video is a basic processing step in diverse biological research [4],[5].

For tracing various intracellular objects in cell imaging, multi-target tracking methods [6],[7] have been exploited previously. However, in the present application domain, tracking-based methods have not been successful for tracing mitochondria for several reasons. First, while all objects are moving in most other cell image sequences, there are many stationary objects in our neural image sequences. Over half of the particles are stationary, and some target objects move at low speeds. Frequent merging and splitting of stationary and moving particles, and sudden starting and stopping of moving targets make it hard to trace individual objects. Second, in the typical input image, the size of a moving target is very small. A target occupies only a few pixels in images. In addition, most particles appear as small dots with similar shapes and brightness. Thus, it is hard to identify individual particle based on shape and brightness information only.

An important characteristic of moving targets is that they are moving only through axons. Thus, once axons are traced first in images, the motion of the target object can be traced by locating curvilinear trajectories. While input video is a collection of images, a kymograph is constructed by combining temporal variation of image intensities on a selected axon. A kymograph is a time-space plot illustrating the intensity changes along an axon as a function of time. It is much easier to trace the curve on a 2D kymograph than finding and tracking small dots on 3D image sequences. Thus, many previous research works on mitochondria tracking have utilized kymographs for analysis. Techniques using image correlation [8] and Hough transform [9] were proposed for the analysis of axonal transports.

An automated kymograph analysis was proposed for tracking secretory granules [10]. As kymograph analysis obtains wider acceptance, automated analysis techniques have been proposed recently [11],[12],[13],[14]. However, since the performance of automated techniques usually depends on the characteristics of input images, the application of these automated techniques to other application domains have been somewhat limited. Recently neural net-based machine learning techniques have been applied to biomedical application domains as well [15],[16],[17],[18]. U-Net architecture has been successful in this application area [15]. An internet-based kymograph analysis tool [19] has been proposed using U-Net architecture.

In this paper, an integrated procedure is described for tracking moving objects trajectories from image sequences acquired from a microfluidic culture platform. The proposed approach is based on a kymograph analysis. Since the particles move along the axons in this application area, axon structures need to be detected first from the input image sequence. This process has been typically performed using curve trace techniques [20],[10]. In our approach, we apply kymograph analysis technique to detect axon structures. Kymograph is an image on 2D time-space domain. While the input is a 2D image defined on (x, y) plane, by regarding the vertical direction as the time axis, vertically oriented axons can be detected by a kymograph analysis process. Similarly, horizontally oriented axons can be detected by applying the kymograph analyzer after rotating the image by 90 degree. Once axons are detected from input image sequences, multiple kymographs are generated along the detected axons.

Using multiple kymographs generated from the image sequence, a trajectory detection process is applied to kymograph data set. Finally, an integration process is applied to each horizontal and vertical kymograph data set to obtain the particle motion information from the image sequence. We have applied the proposed technique to image sequences in our application area. Experimental results will be presented using time-lapse image sequence data.

II. Kymograph Analysis

In biomedical image applications, tracking particles on complex trajectories has been one of the basic processing tasks [4],[6],[7]. In practical applications, low image quality usually makes it impractical to track and analyze particle movements directly in images. While some methodologies have been proposed to detect particle movements directly on images, applications to other domains have been limited.

In time-lapse image sequences obtained from microfluidic culture platforms, particle motions usually arise on axons only, which remain as stationary curves in images. It is often unnecessary to track particles on 3D space. A kymograph is a 2D image depicting the temporal variations of image intensities along an axon curve. Since it is much easier to trace motions of target objects in kymographs than in video frames, kymographs have been utilized as intermediate target images for object tracking. A number of methods have been proposed for enabling kymograph analysis, and some methods have provided software packages for public access [12],[13],[14]. Most of previous methods have somewhat limited applicability, depending on the application domains and program usability.

As deep learning techniques have become successful in image recognition and segmentation areas, machine learning approaches have been adopted to biomedical applications. Currently, U-Net architecture has been the most successful for biomedical image analysis [15]. A U-Net based architecture, KymoButler [19] has been proposed for kymograph analysis. This architecture has a public-accessible implementation, providing the analysis results from a kymograph image supplied through internet.

III. Methods

A. Kymograph Detection

In microfluidic image sequences, object particles appear as scattered dots on each image frame. Experimental microfluidic image sequences consist of 100 images, taken at fixed time intervals. The purpose of the analysis is to identify moving trajectories and detect lengths and speeds of such motions. Fig. 1 illustrates the analysis procedure and the intermediate sample images on each step. Fig. 1-A depicts the first image of a 100-frame image sequence. At each image, small dark dots are object particles, and the analysis is performed to locate trajectories of moving particles. It can be seen that axon structures are not revealed on the first image. Since a moving particle occupies different positions at different frames, when the image sequence is summed through time, trajectories of moving particles tend to become continuous curves. A kymograph is constructed by projecting the image sequence to either (x, t) or (y, t) plane. In order to derive axons, a consolidated image is composed using the entire image sequence, by taking the minimum value in all image frames at each pixel. Here, since moving mitochondria appear on different locations, axons tend to become apparent. Fig. 1-B shows the consolidated image for Fig. 1-A, and it can be seen that each motion trajectory appears as a continuous curve.

Since there are many moving particles on each image sequence, a kymograph is generated separately for each axon curve. From the consolidated image in Fig. 1, it can be seen that several continuous axons have become apparent. After an axon detection algorithm is applied to the consolidated image, eleven continuous axons are detected, as shown in Fig. 1-C as the red or blue curves. A kymograph analysis tool [19] has been applied for the axon detection. Like other kymograph analysis tools, this tool tries to detect a continuous curve along the vertical (time) axis. We apply the tool twice, once to the original image, and once to the 90° rotated image. In Fig. 1-C, the red and the blue curves represent the axons detected by the two trials applied to the original and the rotated images.

In general, there may be many moving particles on each axon. In order to detect all moving particles on the image sequence, each axon needs to be examined separately.

Fig. 1-D and E are the two kymographs generated along the axons marked by small red and blue triangles on Fig. 1-C. Each kymograph in Fig. 1-D was generated by composing image intensities along the specified axon. Each horizontal scanline on the kymograph depicts the image intensity on the axon at each time step. Since the orientations of the red and the blue triangle axons are near horizontal or vertical, Fig. 1-D and E have been obtained by projecting the image sequence to the horizontal and vertical directions. On kymograph, a vertical, near-straight line represents the image of a stationary particle, and a slanted curve shows a trajectory of a moving particle. To detect the direction and the speed of moving particle, it is necessary to track slanted curves.

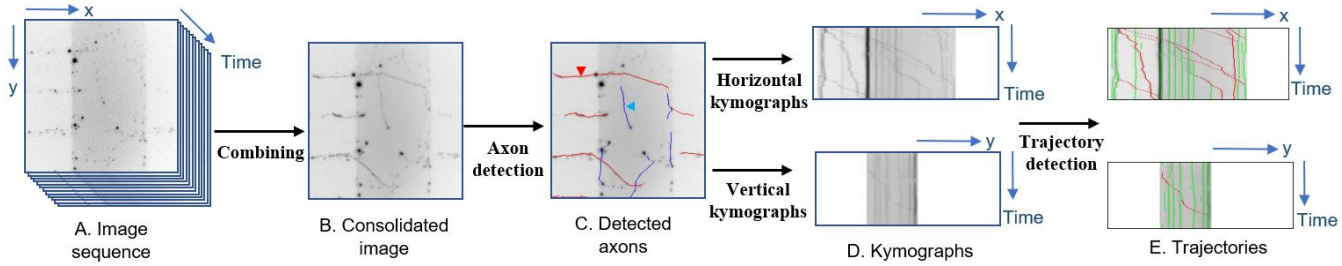


Fig. 1. Image Processing Sequence with the Intermediate Results for an Experimental Neural Image Sequence. The Green and the Red Curves at the Rightmost Column Images denote the Stationary and the Moving Trajectories, respectively.

Two figures in Fig. 1-E show the trajectories obtained from images in Fig. 1-D, detected by the kymograph analysis. After applying the kymograph analysis, each trajectory is classified into stationary and moving curves, as described in Section III-C.

B. Analysis Procedure

The analysis procedure is depicted in Fig. 2. After the consolidated image is constructed from the image sequence, continuous axons are detected using the kymograph analysis process. Axons are then classified into horizontal and vertical groups to decide the direction of the kymograph projection. Depending on the axon orientation, kymograph projection is performed either on (x,t) or (y,t) plane. The number of detected kymographs varies on each axon image. Each kymograph is then analyzed separately, and consolidated trajectories are composed by combining trajectories from all kymographs.

A kymograph image is analyzed using a kymograph analysis tool [19], which can be accessed by supplying each kymograph image through the internet. The result of the analysis is given by a set of points $(s_{k,i}, t_{k,i}), i = 1, \dots, N_k, k = 1, \dots, K$, where K and N_k denote the number of trajectories and the number of points on the k-th trajectory, respectively.

C. Detection of Moving Trajectories

As can be seen in Fig. 1-E, detected trajectories consist of stationary and moving curves. Since the purpose of the kymograph analysis is to extract motion information of moving particles, stationary trajectories are not examined and they need to be removed in a preliminary stage. First, the motion deviation of a trajectory is defined as the difference between the maximum and the minimum horizontal positions, i.e. Δx or Δy. The local speed at each location can be approximated as $v \approx \frac{\Delta s}{\Delta t}$, where Δs denote the difference of positions between neighboring points. Both the motion deviation and the local speed can be computed easily using the trajectory values.

Using the motion deviation and the local speed, the differences of moving and stationary trajectories can be defined as follows.

- Stationary trajectories have small motion deviations: $|\Delta x| < \tau_D$ or $|\Delta y| < \tau_D$, where τ_D is a small value (such as 5 pixels)
- The local speed of a point on stationary trajectories is small: $|v| < \tau_V$, where τ_V is a small speed value.

D. Integration of Multiple Trajectories

Each kymograph depicts particle motions on a single axon. To find the motion trajectories on the entire image sequence, the results from all kymographs need to be integrated. For instance, there have been 11 axons on Fig. 1-C. Since there are kymographs projected into horizontal and vertical directions, the trajectory integration is performed twice through the horizontal and the vertical directions. The integration of the detected trajectories along two orientations is performed as follows.

- Perform trajectory analysis for each kymograph.
- Integrate trajectories from kymographs along the horizontal and the vertical directions separately.

From the integrated trajectory information, it is straightforward to derive the motion information including the number of moving particles at each time, the speed of each particle, the variation of speeds, etc.

IV. Experimental Results

In this research, experiments have been performed using a set of real image sequences, acquired from a microfluidic culture platform using a confocal microscopy. Each experimental video consists of one-hundred 256x256 images, acquired at fixed time intervals. The purpose of the analysis is to detect the trajectories of moving particles. From the trajectories, motion parameters can be computed including the number of moving particles, the length of motion, the widths of motions, and so on.

Each video was analyzed through the procedure depicted in Fig. 2. A consolidated image was composed from the video to reveal the mitochondria trajectories. Since axons have structures similar to kymographs, we have applied the kymograph analysis software available through the internet, to detect axon structures. An example of detected axon structures is illustrated in Fig. 1-C. Since several axons are usually present in a single consolidated image, it is necessary to generate a separate kymograph for each axon. Moving trajectories are detected from each kymograph. The whole motion information is obtained by combining the motion information obtained from the kymograph analysis applied to each separate axon.

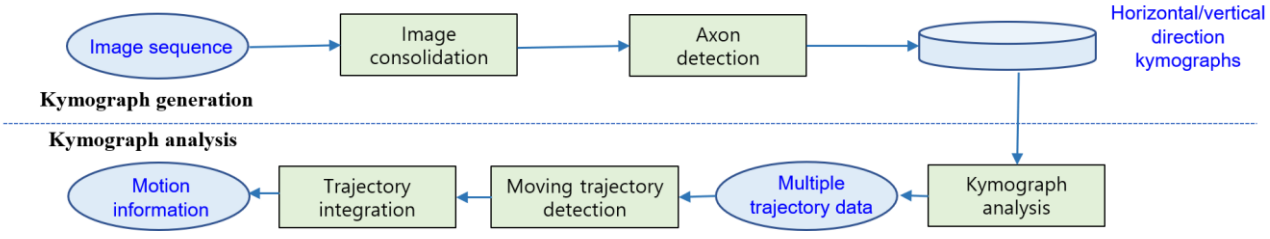


Fig. 2. Proposed Kymograph Analysis Procedure.

A. Accuracy Evaluation of the Trajectory Detection

In order to measure the accuracy of the kymograph analysis technique in this application area, a set of test images were analyzed manually, and ground truth trajectory data were prepared. The performance of the trajectory detection was quantified using the ground truth particle position data prepared manually by tracing the trajectory lines on each kymograph. The detected particle locations were compared with the ground truth positions. A detected point was classified as a true positive match, if there is a corresponding point on the ground data. Since a small amount of deviation may arise in point location, a point on the same time position located within 3-pixel horizontal distance was classified as a matched correspondence. A detected point with no ground correspondence is classified as a false match.

Ground truth data were prepared for four representative kymographs obtained from four different image sequences. Since the complexity of the input image sequences tend to be similar on different image sequences, similar levels of the analysis performance have been observed on other sequence images.

Ground truth and detected trajectories are shown in Fig. 3. Since only motion information is utilized in this research, only moving trajectories are denoted in the ground truth images. expressed as time-position pair. Notice that each trajectory curve looks like a collection of discrete points rather than a continuous curve, since the trajectory consists of a separate point representing each time-position location. The detected points can be seen at the bottom row of Fig. 3.

The accuracy of the trajectory detection is measured using two parameters, Recall and Precision, defined as follows.

$$Recall = \frac{Number\ of\ detected\ points\ on\ ground\ truth}{Number\ of\ ground\ truth\ trajectory\ points} \quad (1)$$

$$Precision = \frac{Number\ of\ detected\ points\ on\ ground\ truth}{Number\ of\ detected\ trajectory\ points} \quad (2)$$

Recall and Precision for the test kymographs are shown on Table I. Since Recall is approximately above 85%, it can be seen that most moving particles are detected correctly.

The results of the kymograph analysis are a sequence of time-position data pair. In Fig. 3, the stationary and the moving trajectories are denoted using different colors. This classification was carried out using the motion detection rules described in Section III-C, that motion trajectories have narrow widths in stationary curves. It can be seen that Precision is lower than Recall. The reason for this phenomenon can be observed by comparing ground truths and detection results in Fig. 3, where it can be seen that some segments are mixtures of stationary and moving parts, while only moving parts are marked in the ground truth data. Since the purpose of the analysis is to detect the moving parts, the value of Recall is more important than that of Precision.

B. Integration of the Detected Kymograph Trajectories

To find the motion trajectories from the entire image sequence, the results from all kymographs detected from multiple axons are integrated. There have been 11 axons on Fig. 1-C. Fig. 4 and 5 depict the detected trajectories from the horizontally (red) and the vertically oriented (blue) axons in Fig. 1-C. Here, moving and stationary trajectories are denoted using the red and the green colors, respectively. It can be seen that moving trajectories have been detected correctly. From the detected trajectories, motion parameters including the amount of movement and the speed can be computed.

TABLE I. DETECTION RATES

Kymograph	***Kymo 1***	***Kymo 2***	***Kymo 3***	***Kymo 4***
# Detected points on ground truth (A)	378	283	445	129
# Ground truth points (B)	405	331	447	142
# Detected points (C)	481	389	593	222
Recall (A/B)	0.933	0.855	0.996	0.908
Precision (A/C)	0.786	0.728	0.750	0.581

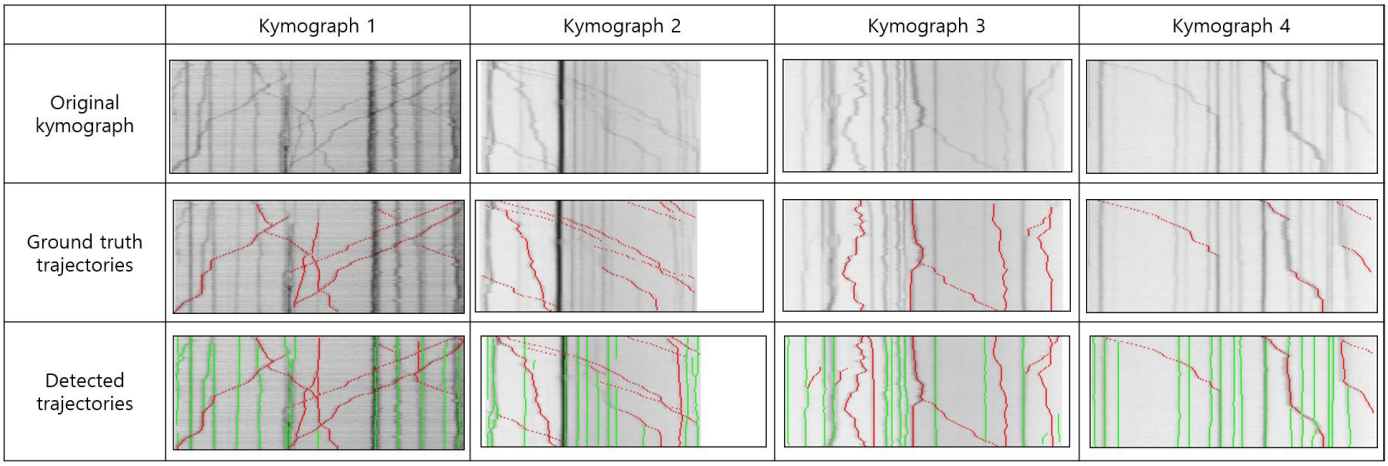


Fig. 3. Comparison of the Detected Trajectories and Ground Truth Data. The Vertical Axis of each Image Denotes the Time Axis.

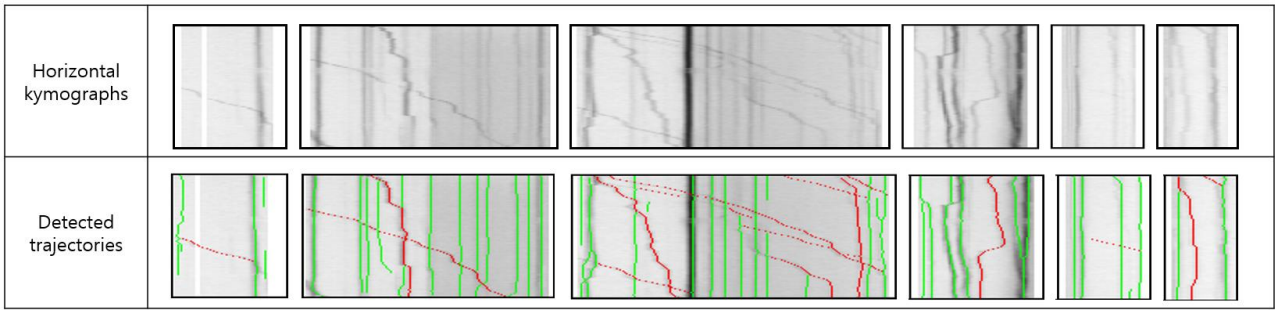


Fig. 4. Results for the Horizontally Oriented Kymographs Generated from the Axons in Fig. 1.

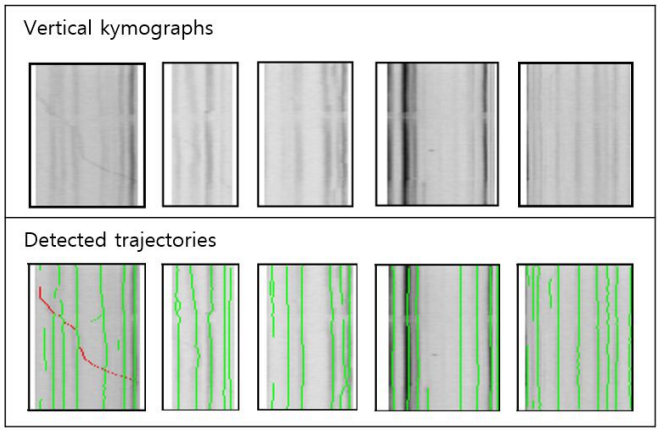


Fig. 5. Results for the Vertically Oriented Kymographs Generated from the Axons in Fig. 1.

V. Conclusion

An integrated procedure has been described for tracking moving object trajectories from image sequences acquired from a microfluidic culture platform. The proposed approach is based on the kymograph analysis. Since particles move along the axons in this application, the axon structures need to be detected first from the input image sequence. We apply a kymograph analysis technique to detect axon structures from the consolidated image of the input image sequence. While the input is a 2D planar image, by regarding the vertical direction as the time axis, vertically oriented axons can be detected by a kymograph analysis process. Similarly, horizontally oriented axons can be detected by applying the kymograph analyzer after rotating the image by 90 degree. Once axons are found from input image sequences, multiple kymographs are generated along the detected axons.

A trajectory detection process is then applied to each kymograph data set. To obtain the particle motion information from the entire image sequence, the integration process is applied to each horizontal and vertical kymograph data set. The proposed technique is applied to image sequences in our application area. It has been demonstrated that practical results can be obtained using time-lapse image sequence data, where the detection accuracy is comparable to other kymographic analysis.

Acknowledgment

This research was supported by Hankuk University of Foreign Studies Research Fund.

References

[1] K. E. Miller and M. P. Sheetz, "Axonal mitochondrial transport and potential are correlated," J. of Cell Science, vol. 117, pp. 2791–2804, 2004.

[2] A. Taylor, M. Blurton-Jones, S. Rhee, D. Cribbs, C. Cotman, and N. Jeon, "A microfluidic culture platform for CNS axonal injury, regeneration, and transport," Nature Methods, vol. 2, pp. 599-605, 2005.

[3] J. Park, B. Vahidi, A. Taylor, S. Rhee, and N. Jeon, "Microfluidic culture platform for neuroscience research," Nature Protocols, vol. 1, pp. 2128-2136, 2006.

[4] I. F. Sbalzarini and P. Koumoutsakos, "Feature point tracking and trajectory analysis for video imaging in cell biology," J. of Structural Biology, vol. 151, pp. 182-195, 2005.

[5] Y. Kalaidzids, "Intracellular objects tracking," Euro. J. of Cell Biology, vol. 86, pp. 569-578, 2007.

[6] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser, "Robust single particle tracking in live cell time-lapse sequences," Nature Methods, vol. 5, pp. 695-702, 2008.

[7] I. Smal, K. Draegestein, N. Galjart, W. Niessen, and E. Meijering, "Particle filtering for multiple object tracking in dynamic fluorescence microscopy images: Application to microtubule growth analysis," IEEE Trans. Med. Imag., vol. 27, pp. 789-804, 2008.

[8] O. Welzel, D. Boening, A. Stroebel, U. Reulbach, J. Klingauf, J. Kornhuber, and T. Groemer, "Determination of axonal transport velocities via image cross- and autocorrelation," Eur. Biophys. J., vol. 38, pp. 883-889, 2009.

[9] O. Welzel, J. Knorr, A. Stroebel, J. Kornhuber, and T. Groemer, "A fast and robust method for automated analysis of axonal transport," Eur. Biophys. J., vol. 40, pp. 1061-1069, 2011.

[10] A. Mukherjee, B. Jenkins, C. Fanf, R. J. Radke, G. Banker, and B. Roysam, "Automated kymograph analysis for profiling axonal transport of secretory granules," Medical Image Analysis, vol. 15, pp. 354-367, 2011.

[11] N. Chenouard, J. Buisson, I. Bloch, P. Bastin, and J-C Olivo-Marin, "Curvelet analysis of kymograph for tracking bi-directional particles in fluorescence microscopy images," Proc. IEEE int. Conf. Image Processing, pp. 3657-3660, 2010.

[12] S. Neumann, R. Chassefeyre, G. E. Campbell, and S. E. Encalada, "KymoAnalyzer: a software tool for the quantitative analysis of intracellular transport in neurons," Traffic, vol. 18, pp. 71-88, 2017.

[13] K. Chiba, Y. Shimada, M. Kinjo, T. Suzuki, and S. Uchida, "Simple and direct assembly of kymographs from movies using Kymomaker," Traffic, vol. 15, pp. 1-11, 2014.

[14] P. Mangeol, B. Prevo, and E. J. G. Peterman, "KymographClear and KymographDirect: two tools for the automated quantitative analysis of molecular and cellular dynamics using kymographs," Molecular Biology of the Cell, vol. 27, pp. 1948-1957, 2016.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Comvolutional networks for biomedical image segmentation, MICCAI, pp. 234-241, 2015.

[16] X. Cai, X. Li, N. Razmjooy, and N. Ghadimi, "Breast cancer diagnosis by convolutional neural network and advanced thermal exchange optimization algorithm," Computational and Mathematical Methods in Medicine, 2021.

[17] Z. Guo, L. Xu, Y. Si, and N. Razsmjooy, "Novel computer-aided lung cancer detection based on convolutional neural network-based and feature-based classifiers using metaheuristics," Int. J. Imaging Syst. Technol., vol. 31, pp. 1954-1969, 2021.

[18] A. Hu and N. Razmjooy, "Brain tumor diagnosis based on metaheuristics and deep learning," Int. J. Imaging Syst. Technol., vol. 31, pp. 657-669, 2021.

[19] M. Jakobs, A. Dimitracopoulos, and K. Franze, "KymoButler, a deep learning software for automated kymograph analysis," eLife, pp. 1-19, 2019.

[20] E. Meijering, M. Jacob, J.-C. F. Sarria, P. Steiner, H. Hirling, and M. Unser, "Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images," Cytometry Part A, vol. 58A, pp. 167-176, 2004.

LPRNet: A Novel Approach for Novelty Detection in Networking Packets

Anshumaan Chauhan, Ayushi Agarwal, Angel Arul Jothi, Sangili Vadivel
Department of Computer Science, Birla Institute of Technology and Science
Pilani, Dubai Campus, Dubai, United Arab Emirates

***Abstract*—Novelty Detection is a task of recognition of abnormal data points within a given system. Recently, this task has been performed using Deep Learning Autoencoders, but they face several drawbacks which include the problem of identity mapping, adversarial perturbations and optimization algorithms. In this paper, we have proposed a novel approach LPRNet, a Denoising Autoencoder which uses algorithms such as Least Trimmed Square, Projected Gradient Descent and Robust Principal Component Analysis, to solve the above-mentioned problems. LRPNet is then trained and tested on NSL-KDD dataset, and experiments have been performed using Accuracy as performance metric for comparing the existing models with the proposed model. The results show that LRPNet has the maximum accuracy of 95.9% and performed better than all the previous state-of-the-art algorithms.**

Keywords*—Novelty detection; deep learning; autoencoders; unsupervised learning

I. Introduction

Novelty detection is classification of points whose characteristics are different from that of normal data [1]. These points which are not like the normal data are called anomalies. The process of anomaly detection is also known as outlier detection or out of distribution detection. Automatic anomaly detection is a task that is of high demand in the areas of fraud detection, network intrusion detection and several other fields.

Earlier the task of anomaly detection was a binary classification problem, where they used to train machine learning algorithms on the normal data as one class, and all the other data was treated as other class. Data which was categorized in this other class was taken as an anomaly point. Classic machine learning algorithms such as Support Vector Machines, Isolation Forest (also known as iForest), and many more algorithms have been used for the task of binary classification. Novelty detection approaches can be classified into 3 types, Density based classification algorithms, distance-based classification algorithms (also referred as clustering algorithms), and deep learning algorithms. The drawback of classical anomaly detection models was that they were inconsiderate about the temporal nature of data, that is, they classified points based on its value and not on the value of previous data fed to it. Due to this their results were not up-to the mark.

In recent years, there has been many advances in deep learning which has led to models which have performed significantly well in anomaly detection as compared to classical anomaly detection models. Algorithms and models which have shown significant results [2][3] are:

- Deep learning models: Autoencoders, Recurrent Neural Networks (RNNs) along with Long Short-Term Memory (LSTM)
- Dimensionality reduction techniques such as Self-Organized Maps (SOM), Randomized Principal Component Analysis (RPCA)
- State space models

This paper discusses the drawbacks of existing deep learning models used for anomaly detection. Also, this work proposes a methodology which overcomes some of those drawbacks. Deep neural networks follow two learning approaches: supervised or unsupervised. The problem with supervised learning, is that they require enough anomaly data along with the normal data, which is very hard to find, as anomalies do not happen every now and then. So due to this, the model is not trained well and does not give promising results. Therefore, most of the models follow an unsupervised approach.

Few drawbacks faced by existing methods [5] are:

- Problem of identity mapping
- Adversarial perturbations
- Optimization algorithms
- Appropriate data.

In this paper, we propose a denoising autoencoder model following an active learning approach, which incorporates projected gradient descent to overcome the drawback of adversarial perturbations and optimization. Robust estimation using Least Trimmed Squares (LTS) is used to prevent the model from adverse effects of outliers on the reconstruction error. After going through a lot of existing literature we found out that 2D-CNN architecture of autoencoder is well suited for this problem, as it has enough layers and uses nonlinear activation functions, which solves our problem of identity mapping.

The paper is organized as follows. Section II comprises a brief working information about the different Autoencoders. In Section III we have provided the issues involved with the currently used Autoencoder methodologies. Section IV comprises the brief description of relevant papers on Novelty Detection. In Section V a description of the dataset being used

is given. Section VI mentions the model development for LRPNet and Section VII comprises proposed methodology. Experiments and results have been given in Section VIII. Finally, in Section IX we conclude the paper along with future scope.

II. Background Theory

In this section we will see about autoencoders and have a brief description of different types of autoencoders.

A. Concept of Autoencoders

In Convolutional Neural Network (CNN), after a convolution layer, the size of input decreases. This decreased output is the features which are useful for solving the problem at hand. But it should be noted that the model learns those features in an unsupervised manner. After the loss of information, we want that information left at hand should be the one crucial for problem solving. From this idea emerged the concept of Autoencoders. Autoencoders is a feed-forward type of artificial neural network, which uses the concept of compression, that is, first the original data is reduced to data of low dimensional space. This job is done by one half of the network called as encoder. The other half does the exact opposite job, it tries to reconstruct the input from this latent space representation, generated by encoder. This part is known as a decoder.

Some of the important properties of Autoencoders are:

- Data-specific: Autoencoders are good at finding coding of only those kinds of data for which it is trained on, i.e., we cannot use autoencoders to compress a geospatial image which is trained on NSL-KDD dataset.
- Lossy: The output of autoencoders may not be exactly the same as the input provided to it. This property is also good for some reasons that it is not just performing identity mapping. Therefore, a reconstruction error is calculated to check how much different is the output from the input.
- Unsupervised: Autoencoders follow an unsupervised learning method, as we do not provide them with the labels, we just provide it with raw data.

Autoencoders are used for many purposes except for just dimensionality reduction, there are many applications where autoencoders were used for classification and generative purposes too. In Fig. 1, the working of an autoencoder is shown.

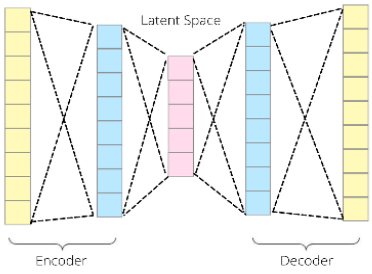


Fig. 1. Working of Autoencoders.

Encoder and decoder part of autoencoders are generally fully connected artificial neural networks and exact mirror images of each other.

B. Briefs of different Types of Autoencoders

Different varieties of autoencoders are used for novelty detection based on the type of input and application. They are as follows:

- Denoising Autoencoder: In these autoencoders, noise (using anomaly data or making some of our inputs as 0) is purposely introduced to prevent the model from simply copying the input to output without learning important features about the data.
- Sparse Autoencoder: Sparsity constraint is known to be a method which leads to good feature extraction. It is done by introducing an extra term to the cost function, which forces the optimization algorithm to reduce the weights of some neurons to almost zero. Due to these reduction in weights, neurons represent their output as a summation of very small values, which ensures the latent space representation to showcase important features (usually the number of neurons in the hidden layer are greater than the number of neurons in the input layer).
- Variational Autoencoder: These are also known as probabilistic autoencoders since their output is sometimes determined by chance even after the training phase is over. They are also known as generative autoencoders, this is because they have capability to generate new outputs that seem to have been taken from training dataset. They are preferred over the Restricted Boltzmann Machine (RBM) as they are easier to train and do faster computation.

Apart from these, there are many other autoencoders such as WTA autoencoders and adversarial autoencoders (but these are not much used in real life application). Amongst all, Denoising autoencoders are the most preferred for novelty detection.

III. Issues Involved in Novelty Detection

There are many challenges related to the input data when given to artificial neural networks (ANN) for training and testing. A combination of numeric and continuous data in the dataset needs to be separated (as a neural network only works with numeric data). Even after separating the different types of data, it is still not ready to give it as an input to the ANN. Some pre-processing steps such as data normalization must be performed so that later when the optimization algorithm changes the weights of the links, it does not change it in a way so that less importance is given to an important feature.

Another important task is to select an appropriate autoencoder. As described in Section II, there are various autoencoders that are used for the job of novelty detection. For novelty detection of cyber-attacks, many autoencoder algorithms such as denoising autoencoder, adversarial autoencoder, convolutional autoencoder and conventional autoencoders have been used. There have been many research papers that review the existing autoencoder architecture based

on the cyber-attack novelty detection, but for different datasets different kinds of autoencoder works fine and others do not.

Issue that arises after deciding on an appropriate autoencoder and doing all the preprocessing steps for the input data is, what will be the architecture of this artificial feed-forward neural network. In architecture of neural network there are many factors involved such as:

- Number of layers for compression (or reconstruction) and respective number of nodes: The number of nodes for the hidden layers should be chosen carefully; it should be in a proper decreasing order (increasing order for decoder), example 128:64:16. We cannot have too few layers as it would make our model underfitting and neither should we have too many as it will make it overfitting. In accordance with the input size, we have used 3 layers for compression (and 3 layers for decompression from the latent space). We have used a brute force method for deciding the number of layers and the number of neurons to be in each layer (shown in Fig. 6). Also, while designing the model we aimed to keep the model as simple as possible, so that it has less computational cost and provides a good accuracy at the same time.
- Different activation functions (or same) for the hidden layers: We can have different activation functions for different layers, but we should keep in mind that the result of these activation functions does not lead us to identity mapping, which is one the most discussed problem of autoencoders. To overcome this problem, in this work, we have used the non-linear activation function Rectified Linear Unit (ReLU), which makes sure there is no instance of identity mapping.
- Optimization algorithm: There are many optimization algorithms that work well, such as Projected Gradient Descent, Adam optimization algorithm, RMSProp, Gradient Descent with momentum and many more. In this work, we have used the Projected Gradient Descent algorithm for the optimization purpose and to prevent adversarial perturbations. Adversarial perturbations are the situation where the autoencoder usually ignores the important data or features that it should learn during compression, and learns the common characteristics that are usually also shared between the normal data and the anomaly data. This mainly leads to the low reconstruction error for normal as well as anomaly data.
- We have used Least Trimmed Squares for the task of Robust Estimation. As Denoising Autoencoders only take the normal data as input, we have to separate out the Outliers from the dataset. In our dataset, when the data points are plotted on a two-dimensional graph, there are some normal data points that are far away from others, and those are removed from the training dataset using LTS.
- Regularization method (to prevent from overfitting): Artificial neural networks often tend to overfit on the training set, to prevent from that as well as for faster convergence we use the concept of regularization. Algorithms such as dropout, L_1 and L_2 regularization, batch norm regularization and many more can be used for regularization. Sometimes one single method is not enough to implement regularization. Therefore, a combination of two methods can also be used. If dropout is not performing well for regularization, we can add batch normalization layers in between the network, and it often tends to solve the problem. In this work, we have used this above-mentioned method for regularization.
- Value of the hyperparameter learning rate is crucial so that the model doesn't have jumps that make it go away from the minimum error point. In this work, we have used 0.01 as the value of learning rate.
- To prevent the model from overfitting, we have used early stopping, which will stop the training phase as soon as the accuracy reaches 95%.

IV. Literature Survey

The problem of novelty detection can be solved using three approaches, as described earlier, that is, density-based approach, distance or clustering based approach and the deep learning-based approach. Density-based approach includes algorithms like GMM, etc. for classifying the anomaly data from the normal data. They use the concept of density of normal data, and make a Gaussian distribution out of it, data which lied in the maximum variance region was classified as anomaly data. OC-SVM, OC-KNN are algorithms which use distance-based approaches for novelty detection. The calculate the distance between A review of all the existing approaches have been given in. In all the experiments, the deep learning approach has performed significantly better than the other two approaches.

Novelty detection is an important task for learning systems in which a subset of the dataset does not fit well on the trained model [1]. Paper concluded that if this data is not in accordance with the data which was used to train the model, then its performance will be affected. In the review, it was mentioned that it is better if we train the model without giving any anomaly data as input and use a statistical approach for classifying the anomaly data. Review of some most used novelty detection techniques was conducted by Dubravko Miljković et al. [2] also concluded the same. After detailing about some important algorithms from each of the 4 approaches (classification-based approach, nearest neighbor-based approach, clustering based approach, statistical based approach) taken for novelty detection, it was concluded that factors such as labelled or n-labeled data, continuous or symbolic features (type of data), and many other factors related to data helps us to decide which will be the most appropriate algorithm for our novelty detection.

A survey of existing outlier techniques, where all different approaches including statistical models (further classified), neural network algorithms, machine learning algorithms and hybrid systems were taken for comparison and conclusion was that models should always be selected based on the dataset. The distribution of dataset, attribute types, and other factors

decide the speed and the accuracy of the model.[3]. It also mentioned that based on whether the data is labelled or not, we decide whether to go for distance-based approach, density-based approach, or novelty approach.

A comparison of different unsupervised anomaly-based approaches used for novelty detection in spatio-temporal data was conducted [4]. They proposed an algorithm that showed better results when the data used for training is scarce as well as having more than 5% of anomaly data. They proposed a hybrid autoencoder based approach which uses convolutional encoder (CAE) along with convolutional Long Short-Term Memory. After testing this model proved to be far better than all the considered methods including iForest, LSTM autoencoder and Convolutional autoencoder.

There are many different possible architectures of autoencoders which can be used for novelty detection. [5] compared some of the best architectures based on computational complexity and accuracy. After experimenting with architectures and other algorithms such as 1D-CNN,2D-CNN, MSCRED, OC-SVM, they concluded that for solving real-time problems, 2D-CNN is the best architecture (showed 100% accuracy in both tests and took minimum time for computation).

Emanuele Principi and Damiano Rossetti et. al [6] evaluated different autoencoder algorithms such as Multi-Layered Perceptron (MLP) autoencoder, Convolutional Neural Network Autoencoder and LSTM for detection of failure of the motor of an electric car. AUC (area under the curve) was the performance metrics, and these methods were trained on 1178 signals (1170 non defective signals and 8 defective signals that were considered as anomaly) and tested on 22 signals (8 normal signals and 14 anomaly signals). Experiments showed that MLP Autoencoder was the best and showed 99.11% accuracy.

Zhiwei Zhnag and Lei Sun [7] proposed an algorithm which uses the concept of along Progressive Knowledge Distillation with Generative Adversarial Networks (GANs), where two different GAN models were combined using the distillation loss. They compared this novel approach with OC-SVM (One Class- Support Vector Machine), Kernel Density Estimation (KDE) and Variational Autoencoder (VAE). Accuracy achieved by the proposed algorithm was 97.8% whereas VAE, KDE and OC-SVM were 96.96%, 81.43% and 95.13% respectively.

Tangqing Li et al. [8] came up with an approach using the concept of re-evaluation of examples after every epoch of training phase has been completed in an autoencoder. They tested this approach on datasets such as MNIST, KDDCUP, and many more, and compared their approach with many baseline models such as OC-SVM and Deep autoencoding Gaussian mixture model (DAGMM). Except for MNIST dataset, where OC-SVM performed better than the proposed algorithm (OC-SVM had an accuracy of 90.2%, Proposed algorithm had an accuracy of 84.2%), for rest all datasets, the proposed algorithm had a better accuracy when compared to all the baseline models.

Stainslav Pidhorskyi and Ranya Almohsen et al. [9] used an adversarial autoencoder with a probabilistic approach for solving the novelty detection problem. They first pre-processed the data by “linearizing the parameterized manifold”, which helps to understand deeply about the normal data (inliers) and then feed it to an adversarial autoencoder. Performance metrics used during experimentations were Area under ROC curve, F1-measure, Area under precision-recall curve and the FPR at 95% TPR (it is the chance that a normal data will be misclassified as anomaly data). The experiment was conducted on MNIST, CIFAR-10, Coil-100 datasets and showed results which were comparable to that of state-of-art algorithms.

The author in [10] explains why autoencoders are a better option for novelty detection than GANs (Generative Adversarial Networks). They stated that GANs during training can face a problem known as mode-collapse, that is, it may map more than 1 input image to a single output image. A full mode collapse situation is rarely encountered, but partial mode collapse can be frequent. Not only mode collapse, GANs are very sensitive to the choice of hyperparameters, non-convergence problems, and many more are the reasons that they are not preferred for novelty detection. Learning of non-semantic features was stated as a problem of autoencoder, that is, it may learn features that share common characteristics between normal and anomaly data, which leads to classification of anomaly data as normal data.

Jorge Meira et al. [11] did a comparative study on the unsupervised anomaly detection techniques used for cyber-attacks. They tested and checked the performance of algorithms such as Autoencoders, Isolation Forest (iForest), One Class-K-means and One Class- Nearest Neighbor on datasets ISCX and NSL-KDD. F1-score, Recall and Accuracy were taken into consideration for comparing the performances of the algorithms. It was noticed that Autoencoder when applied with pre-processing steps such as Z-score and Equal Frequency (EF) showed the best results for both the datasets.

Vishal M. Patel and Pramuditha Perera et al. [12] uses the concept of membership loss function in addition to the mostly used cross entropy error during the training phase of their neural network. For training they also used the knowledge gathered from data apart from what we have in our training dataset to make it learn generic feature filters. When tested and compared performance with VGG16 model on Caltech256 dataset, their proposed model showed superior performance. Accuracy of the proposed model was 93.9% whereas that of VGG16 model was 90.8%.

Autoencoder have shown to perform better when combined with other clustering techniques [13][14]. In [13], autoencoders were combined with. It was experimented on UCSD dataset along with algorithms such as ConvLSTM-AE, Conv2D-AE, Conv-3D AE, and many more. The results clearly showed the supremacy of the proposed algorithms over the others. AUC of the proposed algorithm was 96.5%, whereas the maximum other autoencoders reached was 91.2%. The author [14] used autoencoders with density-based clustering. The latent space encoding and the reconstruction error is sent to a density-based cluster. Points which exceed a

certain error threshold limit are categorized as anomalies. Taking AUC as performance metrics, the model was tested on a range of 20 datasets along with OC-SVM, PCA Based methods and combinations of these methods with density-based clustering. Out of 20 data sets, the proposed method performed better than all in 9 datasets and had the highest average AUC score of 78.29%.

Erik Marchi et al. [15] used Denoising autoencoders with bidirectional LSTM (BLSTM) for acoustic novelty detection. Experiments were conducted on PASCAL CHime speech separation and recognition challenge dataset along with algorithms such as LSTM-CAE (LSTM Convolutional Autoencoder), BLSTM-CAE, Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). The performances were compared on the basis of precision, recall and F1-measure. BLSTM-DAE (proposed algorithm) performed the best amongst all, having a precision of 94.7%, recall of 92.0% and F1-measure of 93.4%.

Yoshinao Ishii and Masaki Takanashi et al. [16] introduced the concept of robust estimation, which not only reduced the computational cost, but also guarantees robustness as its thresholds/restricts the capability of reconstruction of the autoencoder. Group of 15 datasets was used for comparing the performance of the proposed method with normal autoencoder, OC-SVM, iForest and Local outlier factor (LOF). Out of 15, in 7 datasets the AUC score of proposed algorithms was the highest. Proposed algorithm has the highest average AUC score of 85.15.

Chong Zhou and Randy C. Paffenroth et al. [17] stated that Denoising autoencoders are better to use than Maximum correntropy autoencoders, as denoising autoencoders purely trains on the noise free data, due to which its hidden layers are not corrupted (unlike Maximum correntropy autoencoder). They use the concept of RPCA which divides the dataset into two parts (noise free and noise data). Now this noise free data is used for the training of denoising autoencoders. During training they also used L_1 and L_2 regularization techniques as anomaly regularization penalties. The result showed that "the optimal F1-score achieved by iForest was approximately 73% worse than the score achieved by RDA. (Robust Deep Autoencoder)"

Zhaomin Chen and Chai Kiat Yeo et al. [18] evaluated Convolutional Autoencoder, Conventional Autoencoders and Dimensionality based reduction methods on NSL-KDD dataset. It stated that autoencoders are better as, along with performing dimensionality reduction, they also learn the non-linear relationship between the features of the training dataset. The performance metrics was AUC score. For network traffic type UDP, conventional autoencoder performed best, and for rest 3, convolutional autoencoder performed best (not much difference with AUC score of conventional autoencoder).

V. Data Collection

Kaggle is an open-source website, which is used by many machine learning learners and experts for different datasets and participating in various competitions. We are using the NSL-KDD dataset [19] which was provided in the Kaggle Dataset repository for the purpose of novelty detection. NSL-KDD dataset has 2 CSV files, one is used for training purposes and the other for testing purposes. NSL-KDD is a dataset that has 42 features which are shown in Fig. 2.

F#	Feature name	F#	Feature name	F#	Feature name
F1	Duration	F15	Su attempted	F29	Same srv rate
F2	Protocol type	F16	Num root	F30	Diff srv rate
F3	Service	F17	Num file creations	F31	Srv diff host rate
F4	Flag	F18	Num shells	F32	Dst host count
F5	Source bytes	F19	Num access files	F33	Dst host srv count
F6	Destination bytes	F20	Num outbound cmds	F34	Dst host same srv rate
F7	Land	F21	Is host login	F35	Dst host diff srv rate
F8	Wrong fragment	F22	Is guest login	F36	Dst host same src port rate
F9	Urgent	F23	Count	F37	Dst host srv diff host rate
F10	Hot	F24	Srv count	F38	Dst host serror rate
F11	Number failed logins	F25	Serror rate	F39	Dst host srv serror rate
F12	Logged in	F26	Srv serror rate	F40	Dst host rerror rate
F13	Num compromised	F27	Rerror rate	F41	Dst host srv rerror rate
F14	Root shell	F28	Srv rerror rate	F42	Class label

Fig. 2. Features in the Dataset.

Training file is composed of 125973 tuples and the testing file is composed of 22544 tuples. Out of these 148517, 78588 tuples have their label value as 'normal', and the rest all are anomaly packets as shown in Table I. There are a total of 36 other label values that are being treated as anomaly packets.

Our focus is to reject or drop any packet which has even a slight chance of being a malicious packet. Therefore, we are categorizing all 37 labels (types of packets) into broadly two categories; normal and malicious. The traffic proportions of these packets in our dataset are given in Fig. 3.

TABLE I. Frequency of each Label within the Dataset

normal	78588
neptune	47868
satan	4331
ipsweep	4078
portsweep	3302
smurf	3186
nmap	1699
back	1183
warezclient	997
teardrop	996
guess_passwd	464
mscan	310
warezmaster	299
pod	236
apache2	228
processtable	211
snmpguess	99
mailbomb	94
saint	93
buffer_overflow	47
snmpgetattack	43
httptunnel	41
land	20
multihop	16
rootkit	14
loadmodule	13
imap	13
ftp_write	10
ps	9
sendmail	8
phf	5
perl	4
xlock	4
xterm	3
named	2
spy	2
xsnoop	1

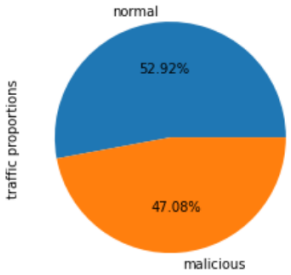


Fig. 3. Traffic Proportions of Dataset.

VI. Theoretical Background on Model Development

In this work, we have used Denoising Autoencoder, Z-score normalization and RPCA. In this section we have provided a brief explanation about these algorithms.

1) Denoising autoencoders: An autoencoder is a special type of deep neural network which is used for the purpose of dimensionality reduction. dimensionality reduction is a process of features selection as well as extraction from a n-featured dataset. Autoencoders are symmetrical in nature, and are basically composed of three components: Encoder, latent space encoding and decoder. The number of layers and the nodes within each layer in encoder and decoder are the same.

Denoising autoencoder is an autoencoder which purposely takes input which has some noise in it, or by making the input corrupt and then do the reconstruction or denoising part. One of the parameters that denoising autoencoder takes is the amount of noise that we want to introduce in the input. The most optimal value used for this parameter is 0.2 provided we have sufficient data. If data is limited then we can also go for higher values. Various applications of denoising autoencoder are feature imputation, anomaly detection, feature extraction and category embedding.

2) Z-score normalization: Z-score is also known as Standard score and it is a technique used to know how far the data point is from the mean of the attribute. More specifically it measures the standard deviation of the data point from the mean of the attribute. It is given by Eq. 1

$$z_i = \frac{x_i - \bar{x}}{s} \quad (1)$$

where x_i is the value of a data point, x bar is the calculated mean of the attribute and s is the standard deviation of the attribute. So, one of the prerequisites to use the z-score is to calculate the mean and standard deviation of the attribute first.

Z-score is used to standardize our dataset to a common range of values so that the autoencoder does not give more importance to a feature which has a high range of values as compared to others. It is one of the most used pre-processing steps for numerical data.

3) Robust principal component analysis: Robust Principal Component analysis is an extension to the most used dimensionality reduction technique Principal Component Analysis (PCA). RPCA is used when we are handling corrupted data or noise data. RPCA returns a low-ranking matrix L_o from a corrupted matrix composed of L_o+S_o. S_o here is a sparse matrix. This decomposition into low-rank matrix and sparse matrix can be achieved by using any of the following techniques: Quantized Principal Component Pursuit method (PCP), Local PCP or Stable PCP. Working of RPCA is shown in Fig. 4.

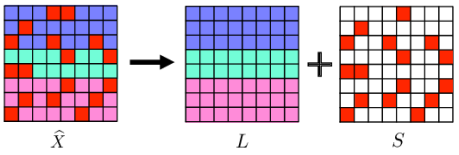


Fig. 4. Working of RPCA [20].

RPCA is used in anomaly detection, face detection, video surveillance and many more applications.

4) Projected gradient descent: It can be considered as a stricter version of Gradient Descent algorithm. In the Gradient Descent algorithm, we use Equation 2 for changing the value of weights and bias of a Neural Network

$$min\, f(\theta)\ \Theta(\text{new}) = \theta(old) - \alpha \Delta J(\theta) \quad (2)$$

Where, α is the learning rate, ΔJ(Θ) is the error (difference between the predicted outcome and the actual outcome) and Θ is the weight of neurons.

We can see that the error is being minimized by moving in a negative gradient direction. In Projected Gradient descent there is a constraint in this equation. It minimizes the error by moving in a negative gradient direction and then project that value onto a valid meaningful set, say *C*. By doing this we make the algorithm more general [21][22].

VII. Proposed Methodology

The steps in the proposed method are explained in this section.

Step 1: Load the dataset D and convert it into a binary dataset, one class being 'normal' and combining all other classes into 'malicious' class.

Step 2: Split the dataset into numerical and categorical column lists based on the datatype (do not include labels in the categorical columns).

Step 3: Remove column 'num_outbound_cmds' from the numerical column list.

Step 4: Encode the labels using Label Encoder. (1- normal and 0-malicious).

Step 5: Normalize the values in all the numerical columns using z-normalization.

Step 6: Form 2 datasets Dnormal and Dattack based on the label values. Use LTS for transferring some data points from Dnormal to Dattac that are far away from other normal data points in a 2-D graph representation.

Step 7: Convert the categorical data into numeric type using get dummies function of Pandas library (does one hot encoding) and combine them with the numerical data.

Step 8: After appending these numerical data in Dnormal and Dattack. Split Dnormal into two: Training (67%) and Testing (33%). Finally, we have the following three datasets: Dnormal-test, Dnormal-train, and Dattackt.

Step 9: Make a copy of all the datasets created in the previous step and process all using Robust Principal Component Analysis. This results in the following three datasets: Dnormal-test-RPCA, Dnormal-train-RPCA and Dattack-RPCA.

Step 10: Create 2 instances of the Proposed Model named as Model 1 and Model 2. Train Model 2 (LRPNet) on the Dnormal-train-RPCA and validate the model on Dnormal-test-RPCA.

Train Model 1, on Dnormal-train and validate the model on Dnormal-test.

Step 11: Test Model 2 on dataset Dattack-RPCA and Model 1 on dataset Dattack using different threshold values and compare them on their reconstruction error score.

The various steps of the proposed methodology are shown in Fig. 5.

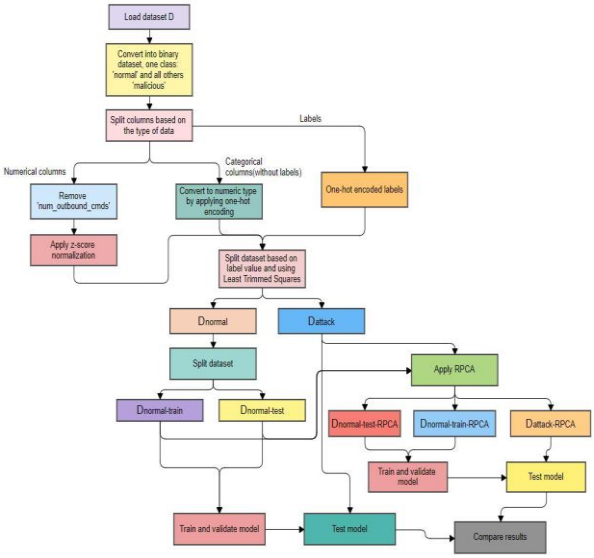


Fig. 5. Proposed Methodology.

VIII. Result and Discussion

We have used Python and Google Collaboratory for experimentation of our model on the NS- KDD dataset. The architecture of the autoencoder used is shown in Fig. 6. Early stopping is used to prevent the model from overfitting.

The results obtained by training and testing the model (Model 1) on the dataset without applying RPCA is shown in Fig. 7 and 8.

```
Model: "sequential_19"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_103 (Dense)            (None, 96)                11712
_________________________________________________________________
dense_104 (Dense)            (None, 64)                6208
_________________________________________________________________
dense_105 (Dense)            (None, 32)                2080
_________________________________________________________________
dense_106 (Dense)            (None, 16)                528
_________________________________________________________________
dense_107 (Dense)            (None, 32)                544
_________________________________________________________________
dense_108 (Dense)            (None, 64)                2112
_________________________________________________________________
dense_109 (Dense)            (None, 96)                6240
_________________________________________________________________
dense_110 (Dense)            (None, 121)               11737
=================================================================
Total params: 41,161
Trainable params: 41,161
Non-trainable params: 0
_________________________________________________________________
```

Fig. 6. Architecture of the Autoencoder.

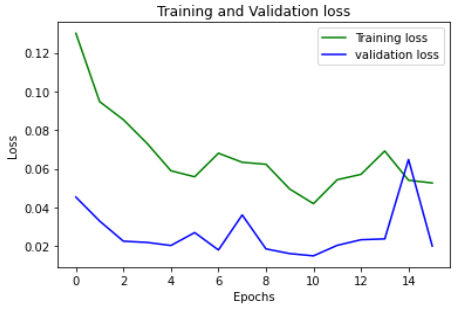


Fig. 7. Loss V/S Epochs Graph of Model 1.

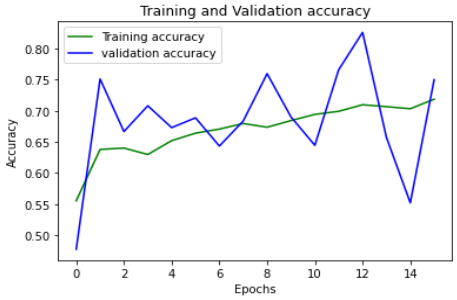


Fig. 8. Accuracy V/S Epochs Graph of Model 1.

The results of training the model (Model 2) on the dataset after applying RPCA is shown in Fig. 9 and 10.

Fig. 9. Loss V/S Epochs Graph of Model 2.

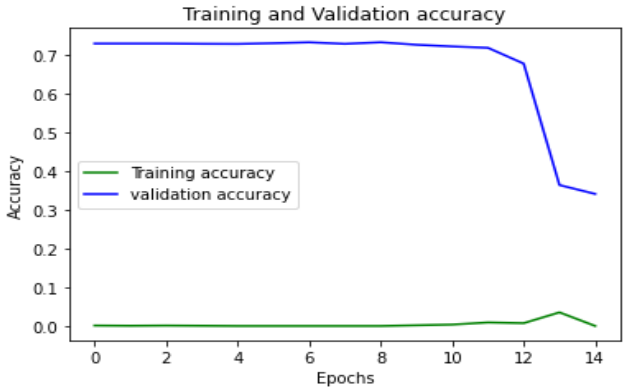


Fig. 10. Accuracy V/S Epochs Graph of Model 2.

We compared the accuracy of the models when RPCA is applied and when it is not applied. For Model 2 when RPCA was applied the reconstruction error for the normal data was high, around 33% and the reconstruction error of malicious packets was around 56%. In this case deciding a threshold value is very difficult because it has to be selected using a brute force method and selecting the one which corresponds to the maximum accuracy, but the prediction accuracies are very high. In Model 1, where RPCA was not applied before training the model, the reconstruction error for normal data was around 14% and for malicious packets was 55%, in this case the difference is quite large, and we can choose an appropriate value accordingly.

For different values of threshold, we got different levels of accuracies for both models which are formulated in Tables II and III.

TABLE II. ACCURACY RESULTS FOR DIFFERENT THRESHOLD VALUES FOR MODEL 2 (LPRNET)

Threshold value	Accuracy (%)
0.33	95.9%
0.35	91.4%
0.37	86.19%
0.39	79.2%
0.41	69.9%
0.43	46%

TABLE III. ACCURACY RESULTS FOR DIFFERENT THRESHOLD VALUES FOR MODEL 1

Threshold value	Accuracy (%)
0.12	95.8%
0.14	94%
0.16	90.91%
0.18	87.27%
0.20	85.76%

TABLE IV. COMPARISON OF MAXIMUM ACCURACY ACHIEVED BY EXISTING METHODS: COMPRESSION AUTOENCODER WITH BLSTM (BLSTM-CAE), COMPRESSION AUTOENCODER WITH LSTM (LSTM-CAE), DENOISING AUTOENCODER WITH BLSTM (BLSTM-DAE) AND DENOISING AUTOENCODER WITH LSTM (LSTEM-DAE)

Model	Accuracy (%)
LSTM-CAE	91.5%
BLSTM-CAE	92.7%
LSTM-DAE	93.4%
BLSTM-DAE	93.6%
LPRNet	95.9%

It could be noted from Tables II and III that the accuracy scores are better for both the models. Though RPCA has shown great results in other fields such as face recognition and video surveillance, in this work RPCA is not useful to a great extent as it is an overhead and the difference in accuracy is also not very high. Another disadvantage that we encountered was, with slight change in the value of threshold, the accuracy of the model depreciated at a very fast rate. Unlike Model 2, Model 1 (RPCA not applied) has a good accuracy over a range of threshold values. In Table IV comparative results with the existing architectures have been formulated and the proposed methodology shows the best results amongst all. The denoising autoencoder architecture proposed is efficient as it takes less than a minute in training and has shown great results provided the applicant choses proper threshold value for classifying between normal and malicious packets.

IX. CONCLUSION

This paper surveys different types of autoencoders that are used for novelty detection and states the issues that are involved while using autoencoders. The survey depicts that denoising autoencoders is the best approach for novelty detection and can have high performance if combined with techniques such as RPCA, clustering based methods and other tools. It also reveals that a proper architecture such as 2D-CNN should be used. Moreover, it could also be concluded that conventional autoencoders are not efficient for novelty detection until used with LSTM. Convolutional and LSTM Autoencoders have better performance but also have a drawback of high computational cost, and in real-time novelty detection, time is an important factor of consideration, therefore they are not preferred over denoising autoencoders. In this work, we experimented with the proposed denoising autoencoder model on the NSL-KDD dataset after applying proper pre-processing for novelty detection purposes. The results showed that the proposed denoising autoencoder achieved a maximum accuracy of 95.9%. The training is also not time consuming, and the accuracy achieved also shows high accuracy scores.

REFERENCES

[1] Marsland, Stephen. "Novelty detection in learning systems." *Neural computing surveys* 3.2 (2003): 157-195

[2] Miljković, Dubravko. "Review of novelty detection methods." *The 33rd International Convention MIPRO*. IEEE, 2010.

[3] Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." *Artificial intelligence review* 22.2 (2004): 85-126.

[4] Karadayi, Yildiz, Mehmet N. Aydin, and Arif Selçuk Öğrencí. "Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data Using Deep Learning: Early Detection of COVID-19 Outbreak in Italy." *IEEE Access* 8 (2020): 164155-164177.

[5] Meire, Maarten, and Peter Karsmakers. "Comparison of deep autoencoder architectures for real-time acoustic based anomaly detection in assets." *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. Vol. 2. IEEE, 2019.

[6] Principi, Emanuele, et al. "Unsupervised electric motor fault detection by using deep autoencoders." *IEEE/CAA Journal of Automatica Sinica* 6.2 (2019): 441-451.

[7] Zhang, Zhiwei, Shifeng Chen, and Lei Sun. "P-kdgan: Progressive knowledge distillation with gans for one-class novelty detection." *arXiv preprint arXiv:2007.06963* (2020).

[8] Li, Tangqing, et al. "Deep Unsupervised Anomaly Detection." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.

[9] Pidhorskyi, Stanislav, et al. "Generative probabilistic novelty detection with adversarial autoencoders." *arXiv preprint arXiv:1807.02588* (2018).

[10] Salehi, Mohammadreza, et al. "Arae: Adversarially robust training of autoencoders improves novelty detection." *arXiv preprint arXiv:2003.05669* (2020).

[11] Meira, Jorge, et al. "Comparative Results with Unsupervised Techniques in Cyber Attack Novelty Detection." *International Symposium on Ambient Intelligence*. Springer, Cham, 2018.

[12] Perera, Pramuditha, and Vishal M. Patel. "Deep transfer learning for multiple class novelty detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[13] Chang, Yunpeng, et al. "Clustering Driven Deep Autoencoder for Video Anomaly Detection." *European Conference on Computer Vision*. Springer, Cham, 2020.

[14] Amarbayasgalan, Tsatsral, Bilguun Jargalsaikhan, and Keun Ho Ryu. "Unsupervised novelty detection using deep autoencoders with density based clustering." *Applied Sciences* 8.9 (2018): 1468.

[15] Marchi, Erik, et al. "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks." *Proceedings 40th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015*. 2015.

[16] Chang, Yunpeng, et al. "Clustering Driven Deep Autoencoder for Video Anomaly Detection." *European Conference on Computer Vision*. Springer, Cham, 2020.

[17] Zhou, Chong, and Randy C. Paffenroth. "Anomaly detection with robust deep autoencoders." *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017.

[18] Chen, Zhaomin, et al. "Autoencoder-based network anomaly detection." *2018 Wireless Telecommunications Symposium (WTS)*. IEEE, 2018.

[19] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," *Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 2009.

[20] Arrigoni, Federica & Fusiello, Andrea & Rossi, Beatrice & Fragneto, Pasqualina. (2015). Robust Rotation Synchronization via Low-rank and Sparse Matrix Decomposition. Computer Vision and Image Understanding. 174. 10.1016/j.cviu.2018.08.001.

[21] Gupta, Harshit, et al. "CNN-based projected gradient descent for consistent CT image reconstruction." *IEEE transactions on medical imaging* 37.6 (2018): 1440-1453.

[22] Bolduc, E., Knee, G.C., Gauger, E.M. *et al.* Projected gradient descent algorithms for quantum state tomography. *npj Quantum Inf* 3, 44 (2017).

A Proposed Model for Improving the Performance of Knowledge Bases in Real-World Applications by Extracting Semantic Information

Abdelrahman Elsharif Karrar
College of Computer Science and Engineering
Taibah University, Medina, Saudi Arabia

***Abstract*—Knowledge Bases are information resources that convert factual knowledge to machine-readable formats to allow users to extract their desired data from multiple sources. The objective of knowledge base population frameworks is to extend KBs with semantic information to solve fundamental artificial intelligence problems such as understanding human knowledge. Information extraction entails the discovery of critical knowledge facts from unstructured text, which is important in the population of knowledge bases. The objective of this paper is to explore the concept of information extraction as a technique for accelerating the performance of knowledge bases with minimal annotation efforts for real-world applications such as content recommendation during a web search. This entails performing slot filling operations for data collection from large KBs and applying probabilistic estimations to determine the accuracy of the new information. The results are then used to explore the feasibility of applying knowledge bases to real-world tasks such as user-centric information access by encoding entities with deep semantic knowledge.**

Keywords*—Semantic information extraction; knowledge base; slot filling; content recommendation

I. INTRODUCTION

Knowledge Base (KB) refers to a specially designed resource for gathering and processing knowledge in logical statement formats that define the relationship between graphical entities. Knowledge Bases utilize a relational knowledge representation framework implemented on artificial intelligence, logic, and semantic networks [1]. Facts representation through KBs follows the guidelines by Resource Description Framework (RDF) in the definition of variable relationships among entities, predicates, and values forming triples such that entities represent people or objects, predicates define entity relationship, and values represent other entities, types, attributes, and values [2]. Triples represent existing facts as illustrated in Table I.

Triples in a knowledge base can be aggregated into a graph composed of directed edges representing relationships and nodes representing values and entities. Edge directions reflect the subject entities in specific triples in the condition of two entities. This implies that edges bridge subject entity to object entity. Different edge types are used to represent various relations through structures known as Knowledge graphs, which enhance the visualization and comprehension of KG structures. [3]

DBpedia is an example of a Knowledge Database, which has been developed by research communities to provide an effective framework for knowledge representation as shown in Fig. 1 [4] [5].

Technology companies such as Google, Microsoft, Facebook, and Yahoo construct and manage in-house Knowledge Bases to perform functions such as answering questions and data querying. The most common knowledge bases operated by technology companies include the Microsoft Graph Satori, Facebook Entity Graph, and Yahoo Knowledge graph illustrated in Table II alongside their relation types, number of entities, and the volume of facts [6].

TABLE I. INSTANCES OF TRIPLES IN KNOWLEDGE BASE

Entity	Predicate	Value
Donald Trump	Age	75
Donald Trump	Profession	Politician, Actor
Donald Trump	Starred in	Apprentice TV show
Apprentice	Genre	Reality Competition
Apprentice	Release Date	January 2004

TABLE II. FEATURES AND ATTRIBUTES OF POPULAR SCHEMATIC KNOWLEDGE BASES

Knowledge Base	Entities	Relation Types	Facts
Google Knowledge Graph	570 million	35,000	18 billion
Yahoo Knowledge Graph	3.4 million	800	1.391 billion
Freebase	40 million	35,000	637 million
DBpedia	4.6 million	1,367	539 million
YAGO2	9.8 million	114	447 million
Wikidata	18 million	1,632	66 million

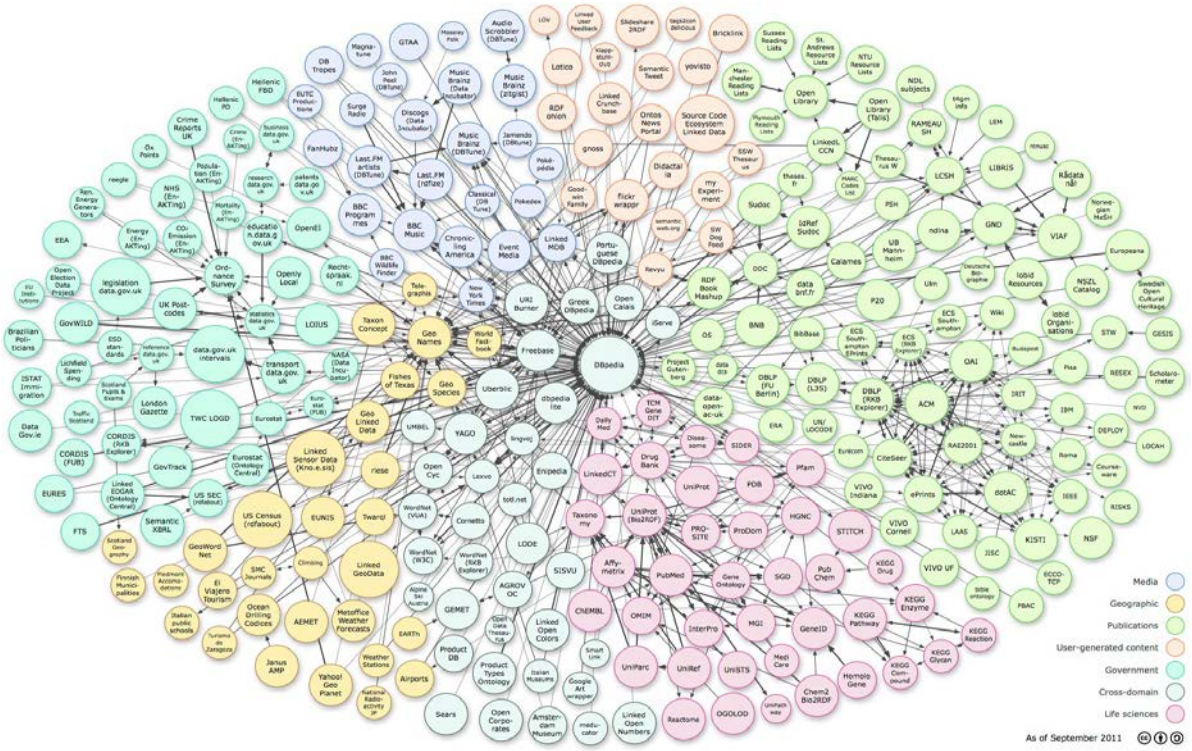


Fig. 1. The Structure of DBpedia Knowledge Base.

Knowledge bases differ from traditional databases in their approach to information management since they are focused on "tables" and "records", which make them efficient when the discovery of new information is not a priority [7]. Knowledge bases are particularly important in domains where the flexibility to link multiple types of information is required. Some of the unique advantages of knowledge bases over traditional data warehouses include:

- Entity-centric: All data is stored based on entity relevance.
- Schema-less: There are no prior requirements for a schema in the knowledge structure.
- Metadata Rich: This contains self-describing metadata streams, which can be easily scaled and integrated across multiple domains.

A. Applications of Knowledge Bases

Knowledge Bases allow for the semantic structuring of computer-readable information, which is a valuable requirement in the construction of intelligent systems [8]. Knowledge bases are a source of power to various big data applications in multiple scientific and commercial domains such as the integration into Google search engine, which stores approximately 0.57 billion entities and 18 billion facts [9]. The Google Knowledge Graph plays an important role in the identification and disambiguation of textual entities to generate enriched search results by semantic structuring of summaries while providing links to related content during explanatory search [10]. Companies typically rely on knowledge bases in gathering information about various entities and their relationships for optimal reuse efficiency in a domain. Knowledge bases are typically used in querying and displaying entity information, recognizing and extracting context, linking entities to data sources and content, discovering and suggesting related information, semantic parsing, and answering questions in technology platforms such as social media and AI-driven virtual assistants.

The role of knowledge bases in utilizing semantic information generated from knowledge graphs to enrich search results is an important milestone towards the transformation of text-based search engines such as Google into semantically-aware question answering platforms. The concept of knowledge graphs has been prominently demonstrated in Watson; a question-answering platform developed by IBM. Watson used a combination of information sources including Freebase, DBpedia, and YAGO to win the game of Jeopardy against a team of human experts [11]. Structured knowledge repositories are integrated into digital assistants such as Amazon Echo, MS Cortana, and Siri by Apple. Knowledge bases such as Freebase store general data generated by its community members from multiple sources including wiki contributions. Knowledge bases have been applied in the Internet Movie Database (IMDb), which is an online storage platform for information related to video games, television programs, and films including character biographies, reviews, crew information, and plot summaries [12].

B. The Concept of Information Extraction

Information Extraction (IE) refers to a process through which structured data is generated from semi-structured or unstructured machine-readable formats [13]. The traditional information extraction systems are used for the efficient

extraction of data from isolated documents using advanced information retrieval methods for data scattered in multiple documents. The systems are capable of identifying the documents containing relevant information and extracting specific facts concerning entities that are conflicting, complementary, or redundant [14].

The first step of data gathering in IE systems is consolidating the known information regarding a specific query entity then searching multiple sources for related information. For instance, if a query 'Donald Trump' is made on an IE system, the objective of slot-filling components is to consolidate information on Donald Trump's place and date of birth, occupation, marital status, education, and any other pre-defined attribute through a process known as 'filling' then adding other related information as recommendations [15]. This process is known as relation extraction since it entails classifying related entities to a relation of interest. For instance, if the system reads a statement 'Donald Trump was born in New York City, the relation born in is extracted to generate search results as (Donald Trump, New York). Information extraction systems are designed to automatically filter information from a pool of sources to fill the missing knowledge base attributes through slot filling before the entities are liked based on their relations.

This research aims at developing a model for improving knowledge basis by extracting information by answering the following research questions;

1) What techniques can be used to construct knowledge bases?

2) How can the accuracy of information extracted from knowledge bases be extracted?

3) In what ways can the efficiency of knowledge bases be improved to perform other tasks such as content recommendation?

This research paper is organized in sections including a review of published literature on the use of knowledge graphs in spoken language understanding, confidence estimation of extracting information systems and the effectiveness of information extraction techniques in improving natural language processing to enrich annotations as well as its role in content recommendation by user profiling in Section II, Section III focuses on the implementation of information extraction techniques and models for improving knowledge bases based on the spoken language understanding (SLU) framework, Section IV explores a high-performance content recommendation model for efficient information extraction from knowledge bases. Section V of this research paper discusses conclusions based on the experimental results and Finally, Section VI provides recommendations for future studies.

II. Literature Review

A. The Population of Knowledge Graphs in Spoken Language Understanding (SLU)

The role of SLU techniques in knowledge bases is to perform slot filling tasks and user intent determination, especially in call routing systems, which are integrated with utterance classification capabilities whereby a speech utterance S_i is categorized into one of M semantic categories, $\hat{C}_r \in C = \{C_{1\ldots}, C_M\}$ given that r represents the utterance index [16]. Researchers have recently developed an advanced slot filling method that involves framing tasks in the form of sequence classification problems to identify the phrase boundaries and labels in a semantic template through deep learning [17] [18]. Slot filling tasks in SLU are defined in the Knowledge Base Population (KBP) whose objective is consolidating information from a large multisource corpus for specific attributes of a query entity. Knowledge graphs are powerful and valuable tools for simplifying research tasks such as computing entity weights to allow the allocation of probabilistic weights in the process of enriching semantic knowledge when detecting SLU relations [19] [20] proposed advanced techniques for processing search queries through semantic parsing in multi-turn dialog systems based on unsupervised natural language processing models.

B. Confidence Estimation in IE Systems

According to [21] confidence estimation refers to a machine learning technique that is used to estimate the confidence scores of a specific output in applications such as machine translation and semi-supervised extraction of relations. The confidence scores of output from speech recognition machines can be computed using a maximum entropy model as described by White and Markov models for singleton tokens.

Another research paper [22] proposed an efficient confidence estimation approach for IE outputs based on machine learning models. This approach worked by computing confidence scores for both multi-field records and extracted fields based on the linear-chain Conditional Random Field (CRF) framework.

However, the machine learning approach is simpler compared to the slot filling technique, which performs complex tasks such as sophisticated inference and co-reference resolution across multiple documents [23]. Inaccurate values extracted in the slot filling operations for KBPs in multiple systems are filtered using techniques such as weighted voting, unsupervised multidimensional truth-finding, heuristic rules, and supervised learning [24].

C. Rich Annotations

Natural Language Processing (NLP) operations such as extracting information can be improved by leveraging user reviews to customize a system to perform personalized searches [25]. Since user reviews may not be readily available, labels created by human annotators, which apply to a range of supervised learning methods can be used to customize the information retrieval system as proposed by [26]. In this case, the traditional machine learning paradigm may be incorporated with a privileged knowledge model to enable the system to accommodate more annotator labels. Recent studies observe an issue with the underutilization of human annotators due to the inclusion of rich annotations into various classification problems [27] [28].

The approach to learning new information through error corrections is conceptualized from the Transformation-based

Error-Driven learning that has been applied to a range of natural language processing operations such as word sense disambiguation, part-of-speech tagging, and semantic role labeling [29]. Rules of transformation in these error-correction techniques are learned automatically based on iteration contexts in each sentence.

D. Content Recommendation by user Profiling

The effectiveness of information extraction techniques for improving Knowledge [30] Bases may be improved through user profiling using factorizing machines and recommendation systems. Research studies suggest that the primary objective of user profiling in IE systems is to align user interests with the recommended items for example in online shopping platforms such as Amazon, the content recommendation in Netflix, or web search customization for enhanced user experience in Google [31].

The functional mechanism of recommendation systems in data extraction may be through content-based recommendation or collaborative filtering, which utilizes matrix factorization and nearest neighborhood techniques to compute user collaboration scores [32].

However, content-based recommendation algorithms work by extracting the unique and dominant attributes that explicitly link users to items, especially in systems with multiple cold start items [33].

According to [34] and [35] researchers have proposed improved approaches for content recommendation by user profiling based on activity ranking, hypergraph learning, latent factor models, probabilistic models, and spatial-temporal model.

A study by [36] observes that developers are now more focused on embedding recommendation and user profiling systems with hierarchical knowledge repositories for the creation of personalized entity recommendations based on knowledge and user activity log obtained from freebase.

A content-based recommendation model proposed by [37] implements a spreading activation algorithm on the DBpedia categorization structure to extract [38]information on user preferences and interests. This technique was later applied to music entity recommendation by Linked Data Semantic Distance (LDSD) with DBpedia by [39] and to movie recommendation by [40]. The recommendation systems are capable of modeling user preferences by exacting information from multiple sources such as implicit and explicit profiles. Deep semantic knowledge provides a framework for extracting rich contextual knowledge of user queries by analyzing the data networks to identify entities in which the users are interested.

The information extraction framework proposed in this paper is consistent with a study [35] which focused on modeling user preferences for customized content recommendation in large knowledge bases primarily relying on data from the Yahoo Knowledge Graph.

III. Methodology

This section focuses on the implementation of information extraction techniques and models for improving knowledge bases based on the SLU framework. Where various knowledge extraction approaches are utilized to identify entities and extract relationships to provide better insights on their application to information extraction based on slot filling and relation detection as the major components of language understanding.

A. Extracting Information from Personal Knowledge Graphs

Rapid technological growth over the past few years has caused a drastic increase in the use of smartphones with advanced capabilities in machine learning, speech recognition, virtual assistants, and voice messaging. Spoken Language Understanding (SLU) features in these information gadgets may be used to extract information from knowledge bases through queries, which may be informational, transactional, or navigational depending on the type of operation being performed. Extracting personal information from Knowledge Bases created by smartphone users may require semantic knowledge graphs due to the high likelihood of data variations [41].

This paper uses schema, a Freebase semantic knowledge graph containing 18 different relations concerning the entity people, person, which may be found in a dataset of spoken utterances. For every relation, a complete set of entities extracted from the Freebase knowledge graph are leveraged in querying the specific entity pairs on the internet using the Bing search engine. The SLU semantic space in this work is aligned to Freebase as a back-end semantic knowledge repository to extract knowledge graph relations in the user utterances as illustrated in Fig. 2.

The user utterances are then classified into binary classes, which may be positive or negative depending on their depiction of personal facts. Once the utterances are formulated as a binary classification problem, the Support Vector Machines (SVM) framework is applied to extract refined factual relations. The *SVM*^{light} package is used to classify the utterances implements binary, linear kernels through a one-vs-rest technique [42]. Identifying the entities and their relations in the utterances, a custom personal knowledge graph for that user is populated with the new information, and the process repeats if the user makes further utterances.

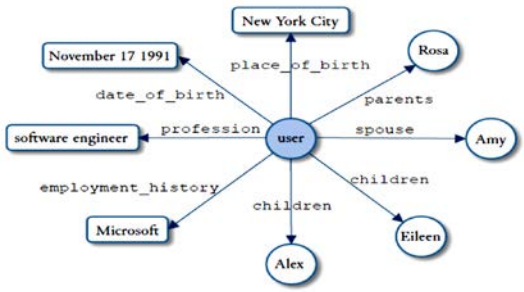


Fig. 2. An Example of a Personal Knowledge Graph.

The training dataset for the framework used in this work is created by searching the internet for related entity pairs in a knowledge graph using the model proposed by [43]. Assuming a web search returns AS as the set containing entity pair a and b, SAS, a subset of AS having

SAS = {s : s ∈ AS ∧ (s, a) ∧ (a, b)} where ∧ (*m*, *n*)is true when *n* is a substring of string *m*. The sentences are then post-processed for the augmentation of relation tags from the knowledge base because some instances may contain multiple relations. For example, if two relations; place of birth (New York, USA) and date of birth (October 17, 1983) about "Brad Hudson" is extracted, post-processing would produce the following instances complete with tags instead of tag-less instances: Brad Hudson was born on <date_of_birth>October 17, 1983, </date_of_birth> in <place_of_birth>New York, USA<\place_of_birth>.

B. Classifying the Personal Assertions

In this experiment, 10 million utterances are extracted from Microsoft KBs and query logs. Factual relations are then mined by extracting personal assertions containing factual relations through the following in the following pattern; 'I am a *, I have a * I live * I was born * I work*'. A random subset of the extracted is selected and annotated whether it satisfies the requirements; it is a personal assertion, invokes relations, and entities can be extracted from the invoked relations. The final dataset contains 12,989 personal assertions out of which only 1,811 utterances contain one or more pre-defined relations. A 10-fold cross-validation technique is then used to create 10 random subsamples whereby 9 subsamples are set aside for training and 1 subsample is retained as a validation set. Cross-validation operations are performed only once on each subsample. From the 236,724 collected samples, 234, 650 are classified accurately (99.12%) and 2,074 are classified inaccurately. This implies that SVM is an efficient classifier for personal assertions.

C. Detecting Relations

The performance of relation detection functionality is determined by testing the models trained using the annotated datasets extracted in the previous section in two scenarios; supervised baseline and unsupervised baseline. A precision model *P@N* was used in the evaluation given that *N* represents positive relations in a given set. From the supervised baseline where 2-fold cross-validation is utilized and the model trained on randomly assigned utterances to two data sets, 84.32% *P@N* upper bound precision is obtained while the unsupervised technique attains 42.85% *P@N* upper bound precision.

D. Slot Filling

The supervised technique was used to perform the slot filling operation due to the variations in semantic annotation mechanisms of the sampled set. The slot F-Measure model was applied to the CoNLL processing script, which attained 68.34% performance efficiency. The model achieves higher performance efficiency when applied to minimal annotations and nontrivial tasks as illustrated in Table III.

TABLE III. PERFORMANCE EFFICIENCY RESULTS FOR SLOT FILLING AND RELATION DETECTION IN DATA EXTRACTION

Relation Type	**Count**	**Relation Detection**		**Slot Filling**		
		Unsupervised	***Supervised***	***Supervised***		
		Precision@Count (%)	***Precision@Count (%)***	***Precision (%)***	***Recall (%)***	***F-Measure (%)***
place_of_birth	8	0.00	0.00	0.00	0.00	0.00
religion	8	0.00	50.00	0.00	0.00	0.00
ethnicity	17	0.00	70.59	100.00	17.65	30.00
employment_history	40	7.50	52.50	50.00	12.50	20.00
nationality	47	0.00	63.83	75.00	82.98	78.79
profission	61	0.00	54.10	50.00	1.64	3.72
gender	63	6.35	82.54	90.91	47.62	62.50
date_of_birth	73	46.58	75.34	56.25	36.99	44.63
places_lived	121	2.48	68.59	69.91	65.29	67.52
sibling_s	248	86.29	90.32	85.92	71.08	77.80
children	260	23.08	87.31	80.92	47.31	59.71
parents	401	19.95	86.78	83.97	65.17	73.39
spouse_s	464	82.11	94.39	86.81	68.10	76.33
Total	1811	42.85	84.32	82.01	58.58	68.34

IV. Content Recommendation by User Profiling

The evolution of the Web has positioned the internet as a crucial player in providing users with access to information from multiple sources. Information overload is one of the greatest challenges of the web hence the need for content recommendation to match user interests. Despite the monumental milestones in the design of recommendation systems, there are significant challenges in availing users of high-quality information. This section explores a high-performance content recommendation model for efficient information extraction from knowledge bases. The core objective of user modeling in this framework is to understand their current preferences and predict future interests in contextual applications such as sports databases. The data used in this experiment is obtained from Yahoo News Streams, which contain information such as the sequence of websites that a user has visited as expressed in the (1) for a typical user u;

$$\mathbf{L}^{u} = \langle w^{u}, w^{u}, \ldots, w^{u}, \ldots, w^{u} \rangle \quad (1)$$

Such that w^u represents the websites visited by user u at time t.

Unstructured information containing attributes such as the user location, language, identity, demographics, timestamps, and click/skip labels. Additionally, the Wikipedia Knowledge Graph was used as a knowledge source for enriching feature space by monitoring evolving sources and wrapping different sources.

A. Modeling user Profiles

A high-level Pipeline algorithm is utilized to model user interests and predict preferences and FastEL software is used for linking entities. A separate entity augmentation algorithm is used to extract entities from user logs then link them to the entities in the Wiki KB. The following code is executed to perform this operation;

Input: A sample user opened document D stored in a Global KG G, which contains relation triples defined by $\sigma = E_a, \rho, E_b$ such that ρ represents a relation predicate for n iterations m maximum augmented entities.

1: Generate initial entities $\boldsymbol{E} = \{e\}$ from $\boldsymbol{D}$
2: **repeat**
3: Augment entities using facts from $\boldsymbol{G}$
4: Re-score interest weights of augmented entities
5: **until** converged or reach $\boldsymbol{n}$ iterations
6: **return** top $\boldsymbol{m}$ augmented entities from the list

Named entities can be extracted from the visited web pages and linked to related Wiki entities based on the user logs. However, the entities may not provide adequate information on user interests hence cannot accurately predict future preferences hence the need to leverage the Yahoo Knowledge Graph to augment the entities into relational facts with a higher degree of accuracy. Once the entities are augmented and retrieved, a decayed interest weight is then assigned to indicate the lowest probability that user interests lie in a particular category.

B. The Framework for Profiling users

According to [44] the user profiling model used for content recommendation in search engines utilizes Factorization Machines (FM) to perform latent factoring and matrix factorization in recommender systems. A latent space for every user is constructed to allow for the differentiation of user preferences in the process of learning from the unstructured dataset. A factorization-machine-based latent factor framework is used to decompose every user profile shared and personalized latent factors. The process of mapping profiles into latent factors is standardized for every user hence making it possible to enrich the information for those with minimal interaction data.

C. Experiment

The experiments are based on a sample of 32.09 billion user logs collected from Yahoo News Stream over one month. The user profiles are evaluated for quality by splitting the dataset into training and testing groups based on event timestamps. Data sets from the first three weeks (23.68 billion events) are used for training while data from the fourth week (8.42 billion events) is used for model testing. For the training dataset, each user profile is ranked and performance evaluated based on ground truth labels, which may be positive or negative.

Inner product values are used between item features and user profiles to generate the ranking scores of each user-item pair. The items are then ranked as positive if they have a higher ranking otherwise negative based on metrics such as the Area under the Curve (AUC), Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP) as defined in the (2), (3) and (4);

$$MAP = 1/m \sum_{i=1}^{m} \frac{\sum_{k=1}^{n_i} P(k)}{n_i} \quad (2)$$

$$MRR = 1/m \sum_{i=1}^{m} \frac{1}{r_i^1} \quad (3)$$

$$MAP = 1/m \sum_{i=1}^{m} \frac{(\sum_j r_i^j) - P_i(P_i+1)/2}{P_i * N_i} \quad (4)$$

Given that $P(k)$ represents precision at k, n_i: user-related links, $u_j r^1$: ranking of the links that were clicked first by user u_i, P_i: the clicked links, and N_i: non-clicked links in the profile for user u_i.

When the number of iterations is adjusted to 1, it achieves about 193% relative and 10% absolute performance improvement in mean average precision; 191% relative and 17% absolute performance improvement in mean reciprocal rank, which is significantly high compared to the baseline system, which obtained 12% relative and 7% absolute performance improvement. Mean average precision computes the average precision scores for listed items while the mean reciprocal rank calculates the inverse position of the initially ranked relevant items. Therefore, both MRR and MAP compute ranking scores for listed items. The area under the curve describes the ratio of false positives and true positives when the threshold parameter is varied suggesting that when

entity ranking and coverage are applied to content recommendation through entity augmentation, it extracts additional related entities enriching the feature space significantly according to [45].

V. Discussion and Conclusion

Technological evolution has led to the rapid adoption of online news platforms as a source of information from a wide range of sources across the globe. Due to the high volume of documents on millions of websites, users face many challenges finding their articles of interest or any other precise information. Knowledge Bases such as Wikipedia are rich information resources for users seeking knowledge in various fields including culture, technology, science, and history. This study sought to improve the efficiency of knowledge bases by analyzing the statistical frameworks for building user-centric KBs and extracting personal facts from user utterances through personal assertion classification.

The study also sought to understand how the accuracy of information extracted from knowledge bases can be validated using a maximum entropy framework. Consequently, a framework for rich annotation-guided learning was developed as an approach for improving the efficiency of knowledge basis through information extraction [13]. The annotation framework was designed with a capability for feature enrichment, which allows for the analysis of relative efficacy and scalability of slot filling operations in KBP settings. A review of previously published studies demonstrates that a slight increase in the annotation period improves KB performance significantly. The study also sought to investigate how knowledge bases can be improved to advance tasks such as content recommendation based on the users' online activity. The experimental findings for these improvement operations in knowledge bases suggest that refining information extraction techniques is an efficient approach to improving the performance of knowledge bases.

VI. Future Work

While researchers have made significant progress towards the understanding of knowledge base architectures, various gaps need to be filled, especially on the categories of knowledge possessed by human beings. Current literature does not provide detailed representations of facts based on common sense and procedural knowledge. Knowledge representation through reasoning and learning remains an important aspect of future studies on the integration of machine learning and artificial intelligence capabilities to information extraction from knowledge bases. Other relevant fields for future research include the population of personal knowledge graphs, confidence estimation for knowledge bases, and guided learning for rich annotations.

References

[1] M. Mutasim and A. Karrar, "Impute Missing Values in R Language using IBK Classification Algorithm," International Journal of Engineering Science and Computing, vol. 11, no. 6, pp. 28328-28338, 2021.

[2] J. Ma, D. Li, Y. Chen, Y. Qiao, H. Zhu and X. Zhang, "A Knowledge Graph Entity Disambiguation Method Based on Entity-Relationship Embedding and Graph Structure Embedding," Computational Intelligence and Neuroscience, vol. 2021, pp. 1-11, 2021.

[3] H. Wu, S. Y. Liu, W. Zheng, Y. Yang and H. Gao, "PaintKG: the painting knowledge graph using bilstm-crf," in 2020 International Conference on Information Science and Education , 2020.

[4] T. P. Tanon, G. Weikum and F. Suchanek, "YAGO 4: A Reason-able Knowledge Base," in European Semantic Web Conference, Lecture Notes in Computer Science, 2020.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in Lecture Notes in Computer Science, 2007.

[6] D. Diefenbach, M. D. Wilde and S. Alipio, "Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph," in The Semantic Web – ISWC 2021. ISWC 2021. Lecture Notes in Computer Science, 2021.

[7] A. E. Karrar, M. A. Abdalrahman and M. M. Ali, "Applying K-Means Clustering Algorithm to Discover Knowledge from Insurance Dataset Using WEKA Tool," The International Journal Of Engineering And Science, vol. 5, no. 10, pp. 35-39, 2016.

[8] A. E. Karrar, "A Novel Approach for Semi Supervised Clustering Algorithm," International Journal of Advanced Trends in Computer Science and Engineering, vol. 6, no. 1, pp. 1-7, 2017.

[9] N. Sahlab, S. Kamm, T. Müller, N. Jazdi and M. Weyrich, "Knowledge Graphs as Enhancers of Intelligent Digital Twins," in 2021 4th IEEE International Conference on Industrial Cyber-Physical Systems, 2021.

[10] M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich, "A Review of Relational Machine Learning for Knowledge Graphs," Proceedings of the IEEE , vol. 104, no. 1, pp. 11-33, 2016.

[11] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer and C. Welty, "Building Watson: An Overview of the DeepQA Project," AI Magazine, vol. 31, no. 3, pp. 59-79, 2010.

[12] Y.-T. Huang and P.-F. Pai, "Using the Least Squares Support Vector Regression to Forecast Movie Sales with Data from Twitter and Movie Databases," Symmetry, vol. 12, no. 4:625, 2020.

[13] A. E. Karrar, "The Use of Case-based Reasoning in a Knowledge-based (Learning) Software Development Organizations," International Journal of Innovative Research in Science, Engineering and Technology, vol. 5, no. 5, 2016.

[14] A. E. Karrar, N. H. Mohammed and M. M. Ali, "Impact of Using Preprocessing in Data Mining and Knowledge Discovery Process," International Journal of Computing and Technology, vol. 3, no. 12, pp. 524-527, 2016.

[15] S. Qiu, P. An, K. Kang, J. Hu, T. Han and M. Rauterberg, "A Review of Data Gathering Methods for Evaluating Socially Assistive Systems," Sensors, vol. 22, no. 1:82, 2022.

[16] X. Sun, J. Gu and H. Sun, "Research progress of zero-shot learning," Applied Intelligence, vol. 51, no. 2, pp. 1-15, 2021.

[17] T. He, X. Xu, Y. Wu, H. Wang and J. Chen, "Multitask Learning with Knowledge Base for Joint Intent Detection and Slot Filling," Applied Sciences, vol. 11, no. 11, 2021.

[18] A. R. Johansen, C. K. Sønderby, S. K. Sønderby and O. Winther, "Deep Recurrent Conditional Random Field Network for Protein Secondary Prediction," in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, Boston, 2017.

[19] F. Orlandi, J. Debattista, I. A. Hassan, C. Conran, M. Latifi, M. Nicholson, F. A. Salim, D. Turner, O. Conlan, D. O'sullivan and J. Tang, "Leveraging Knowledge Graphs of Movies and Their Content for Web-Scale Analysis," in 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2018.

[20] V. Hudeček, O. Dušek and Z. Yu, "Discovering Dialogue Slots with Weak Supervision," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

[21] V. Reshadat, M. Hourali and H. Faili, "Confidence measure estimation for Open Information Extraction," Journal of Information Systems and Telecommunication, vol. 6, no. 1, pp. 1-8, 2018.

[22] A. Raghibi and L. Oubdi, "A Proposed Model for Social Impact Sukuk," TURKISH JOURNAL OF ISLAMIC ECONOMICS, vol. 8, no. 2, pp. 501-516, 2021.

[23] L. Qiu, Y. Ding and L. He, "Recurrent Neural Networks with Pre-trained Language Model Embedding for Slot Filling Task," arXiv preprint, arXiv:1812.05199, 2018.

[24] S. Verlinden, K. Zaporojets, J. Deleu, T. Demeester and C. Develder, "Injecting Knowledge Base Information into End-to-End Joint Entity and Relation Extraction and Coreference Resolution," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021.

[25] C. Orasan and R. Mitkov, "Recent Developments in Natural Language Processing," in The Oxford Handbook of Computational Linguistics 2nd edition, R. Mitkov, Ed., Oxford University Press, 2021.

[26] L. Zhao-Yang and H. Sheng-Jun, "Active Sampling for Open-Set Classification without Initial Annotation," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019.

[27] C. Deng, X. Ji, C. Rainey, J. Zhang and W. Lu, "Integrating Machine Learning with Human Knowledge," iScience, vol. 23, no. 11, pp. 1-27, 2020.

[28] C. David, L. Quentin and D. Alexandre, "Multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies," Scientometrics, vol. 127, pp. 545-575, 2022.

[29] S. Razniewski, A. Yatesa, N. Kassner and G. Weikum, "Language Models As or For Knowledge Bases," arXiv preprint, arXiv:2110.04888, 2021.

[30] A. E. Karrar, "Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer," International Journal of Advanced Computer Science and Applications, vol. 13, no. 1, pp. 181-188, 2022.

[31] M. Nasir, C. I. Ezeife and A. Gidado, "Improving e-commerce product recommendation using semantic context and sequential historical purchases," Social Network Analysis and Mining volume, vol. 11, no. 1, 2021.

[32] Y. Koren, "Factorization meets the neighborhood: a multifaceted collabora- tive filtering model," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008.

[33] A. Al-Bazi, V. Palade, R. A. Hadeethi and A. Abbas, "AN IMPROVED FUZZY KNOWLEDGE-BASED MODEL FOR LONG STAY CONTAINER YARDS," Advances In Industrial Engineering And Management, vol. 10, no. 1, pp. 1-9, 2021.

[34] A. Deepak, C. Bee-Chung, G. Rupesh, H. Joshua, H. Qi, I. Anand, K. Sumanth, M. Yiming, S. Pannagadatta, S. Ajit and Z. Liang, "Activity Ranking in LinkedIn Feed," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.

[35] E. Zhong, L. Nathan, S. Yue and R. Suju, "Building Discriminative User Profiles for Large-Scale Content Recommendation," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[36] S. Manchanda, M. Sharma and G. Karypis, "Distant-Supervised Slot-Filling for E-Commerce Queries," in 2021 IEEE International Conference on Big Data, 2021.

[37] S. K. Cheekula, P. Kapanipathi, D. Doran, P. Jain and A. Sheth, "Entity Recommendations Using Hierarchical Knowledge Bases," in ESWC 2015, 2015.

[38] M. Umair, F. Majeed, M. Shoaib, M. Q. Saleem, M. S. Adrees, A. E. Karrar, S. Khurram, M. Shafiq and J.-G. Choi, "Main Path Analysis to Filter Unbiased Literature," Intelligent Automation & Soft Computing, vol. 32, no. 2, pp. 1179-1194, 2022.

[39] Z. Fattane, K. Mohsen and B. Ebrahim, "User interest prediction over future unobserved topics on social networks," Information Retrieval Journal, vol. 22, pp. 93-128, 2019.

[40] B. Hui, L. Zhang, X. Zhou, X. Wen and Y. Nian, "Personalized recommendation system based on knowledge embedding and historical behavior," Applied Intelligence, vol. 52, pp. 954-966, 2022.

[41] P. Kaur, P. Nand, S. Naseer, A. A. Gardezi, F. Alassery, H. Hamam, O. Cheikhrouhou and M. Shafiq, "Ontology-Based Semantic Search Framework for Disparate Datasets," Intelligent Automation and Soft Computing, vol. 32, no. 3, pp. 1717-1728, 2022.

[42] S. Nurse and J. Bijak, "Building a Knowledge Base for the Model," in Towards Bayesian Model-Based Demography. Methodos Series (Methodological Prospects in the Social Sciences), vol. 17, Springer, Cham, 2022, pp. 51-70.

[43] W. Wu, Z. Zhu, G. Zhang, S. Kang and P. Liu, "A reasoning enhance network for muti-relation question answering," Applied Intelligence, vol. 51, no. 5, p. 4515–4524, 2021.

[44] S.-Y. Jeong and Y.-K. Kim, "Deep Learning-Based Context-Aware Recommender System Considering Contextual Features," Applied Sciences, vol. 12, no. 1, 2022.

[45] C. Chaudhary, P. Goyal, D. N. Prasad and Y.-P. P. Chen, "Enhancing the Quality of Image Tagging Using a Visio-Textual Knowledge Base," IEEE Transactions on Multimedia, vol. 22, no. 4, pp. 897 - 911, 2020.

Melody Difficulty Classification using Frequent Pattern and Inter-Notes Distance Analysis

Pulung Nurtantio Andono*, Edi Noersasongko, Guruh Fajar Shidik
Khafiizh Hastuti, Sudaryanto Sudaryanto, Arry Maulana Syarif
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

***Abstract*—This research proposes a novel method for melody difficulty classification performed using frequent pattern and inter-notes distance analysis. The Apriori algorithm was used to measure the frequency of the notes in the note sequence, in which the melody length is also included in the calculation. In addition, the inter-notes distance analysis was also used to measure the difficulty level of composition based on the distance between successive notes. The classification was performed for traditional Javanese compositions known as Gamelan music. Symbolic representation, in which the Gamelan compositions music sheets were collected as the dataset, was chosen by asking experts to divide the compositions based on their difficulty level into basic, intermediate and advanced classes. Then, the proposed method was implemented to measure the difficulty value of each composition. The difference in the interpretation of the difficulty level between the experts and the difficulty value of the composition is solved by calculating the mean value to obtain the range of difficulty values in each class. Evaluation was performed using confusion matrix to measure the accuracy, precision and recall value, and the results reaching 82%, 82.1% and 82%, respectively.**

***Keywords*—Multi-class classification; frequent analysis; Apriori; Symbolic music; Gamelan**

I. Introduction

This research aims to develop a classifier to discriminate the difficulty level of melody or composition. The topics discussed in this research are closely related to the topic of adaptive learning, including learning to play a musical instrument which becomes the motivation in formulating problems from the difficulty level of composition. Therefore, a brief explanation including the research of adaptive learning literature is deliberately described in this article to emphasize the urgency of the need for a difficulty level classification system in the development of adaptive learning to play musical instruments.

Adaptive learning systems can dynamically adjust learning content based on user abilities and preferences [1]. Adaptive learning systems are developed from simple to complex with a set of rules and self-learning algorithms [2] so that they can provide immediate feedback so that users can stay focused and easily make corrections independently [3]. The difficulty level of the problems that must be solved becomes a challenge in the development of an adaptive learning system. The automatic adjustment of the difficulty level is part of the intelligent tutorial systems that can identify the user's characteristics so that the system can determine the suitable task based on the user's abilities [4]. The automatic adjustment of the difficulty level examples can be found in the works by [5].

Intelligent electronic learning utilizes an adaptive learning approach. However, adaptive learning approaches are rarely found in musical electronic learning, such as in melody learning. Electronic learning to play musical instruments involves melodies, thus classification of melodies (compositions) based on the level of difficulty is still rare as well. Meanwhile, the difficulty level of composition based on the melodic pattern has a positive impact on adaptive learning to play musical instruments [6]. Developing an intelligent musical electronic learning should involve adaptive learning approach. By classifying the difficulty level of composition, the system can determine or provide recommendations containing the composition to be learned based on the user's ability.

Against these facts, a novel approach to develop a classifier that is able to measure the difficulty level of composition was proposed in this research. The classifier was developed using a set of rules defined based on frequent pattern and inter-notes distance analysis. Although melody has a sequence or time series data type, the characteristics of the problems encountered are considered suitable to be solved using frequent mining algorithms rather than sequence mining algorithms. In addition, the inter-notes distance value between successive notes is also used as a parameter in measuring the difficulty level of composition. This makes the analysis of the difficulty level of composition is unique. The forward and backward notes have the same pattern in the inter-notes distance value.

Participation in the preservation of cultural heritage is the motivation in carrying out this research. Thus, the melody difficulty classification system developed in this research is implemented to traditional Javanese compositions known as Gamelan music. However, the results achieved in this research can also be an inspiration to develop a learning system for playing musical instruments of other types of music.

II. Related Work

In electronic learning, an adaptive learning approach is needed to achieve an efficient, effective learning experience and support customization settings for users. Tasks or difficulty classification is part of the adaptive learning that supports the system in determining the suitable task by considering the user's ability. In general, the system will measure the user's answer by assigning an accuracy weight to be used as a basis for determining the level of task that is match the user's

*Corresponding Author

performance. Therefore, questions or tasks to be solved by the user must be classified based on to the difficulty level. In adaptive learning to play musical instruments, classification can be used as a solution in weighting the difficulty level of composition.

Smart devices and artificial intelligence approach are needed to develop a smart learning environment that supports adaptive learning [7]. However, not all electronic learning systems that uses smart devices to run implementing an adaptive learning approach. Moreover, in learning to play musical instruments, the systems generally are limited to the transformation of musical instruments into digital media as in the work of mobile-based gamelan electronic learning media [8], mobile-based hybrid digital-physical harp instrument [9], 3D virtual traditional Chinese instrument called yangqin [10], virtual traditional Brazilian maracatu [11].

A learning system for playing traditional Javanese musical instruments developed by [12] provides a collection of compositions and tempo preferences that are divided into slow to fast ranges based on the time interval between notes switching. However, the collection is not classified according to the difficulty level of composition, and the choice of tempo is determined by the user. In other words, the system has not yet adopted the ability to automatically and dynamically identify user abilities to recommend the suitable composition and tempo.

A task content analysis team was formed to sort the tasks according to their difficulty level [13], while in the similar case; the difficulty level was designed by developing a curriculum [14]. Both of them used English as their learning content, in which the database of questions of various difficulty levels is easier to find than the composition database with groupings based on the complexity of melodic patterns. Meanwhile, analysis of the difficulty level manually performed has a weakness when dealing with a lot of data or adding data to a task. An adaptive learning system to play a musical instrument can be found in the work by [6], in which the system automatically selects the difficulty level of compositions based on the user's true-false count tracking. Unfortunately, the classification method for the difficulty level of composition is not explained.

The Apriori algorithm is a popular algorithm in association relationship analysis or frequent pattern analysis which is also known as market basket analysis. The Apriori algorithm was modified by counting only the sequential transactions that are frequent so that transaction (a, b) is not the same as transaction (b, a). Furthermore, in sequence data mining, many Apriori-like or Apriori-based algorithms have been proposed by researchers, such as mining Web access sequence [15], or modifying the association rules by adding time constraint [16]. Meanwhile, in the melodic pattern, Apriori-based algorithms have been developed by [17] to generate music, and the Apriori based on Function in a Sequence (AFiS) algorithm proposed by [18] counted only the sequential transactions that are frequent with additional procedures in the form of measuring musical elements based on their position in the sequence as functions to identify frequent sequence patterns.

The Apriori algorithm has been used in various classification problems, such as classification of patient care needs by [19], classification in risk prediction to define disaster rules based on models developed using the Neural Networks method [20], admission planning classification and job prediction in education midwives. This combines the Apriori algorithm with the Decision Tree Algorithm [21]. Besides being used without being combined with other algorithms in finding solutions, the Apriori algorithm is usually used to map and analyze data which is then used as a constraint for other algorithms in making decisions, or to define rules based on data obtained from training results using other algorithms. Selection of Apriori algorithm as the main and single algorithm in classification problems is rarely found, especially in time-series data problems. However, this algorithm has the advantage of managing small datasets to construct good classification rules. In this research, the Apriori algorithm is proven to be an alternative in solving classification problems on time-series data.

III. METHODOLOGY

This research aims to measure the difficulty level of composition in order to classified compositions into basic, intermediate and advanced classes. Traditional Javanese compositions were chosen as objects of research with the aim of participating in the preservation of intangible cultural heritage through the implementation of artificial intelligence.

Gamelan music consists of two musical scales, which are pelog and slendro. Each of the musical scale has a different tone frequency. The pelog musical scale consists of seven notes of (1, 2, 3, 4, 5, 6, 7). Meanwhile the slendro musical scale consists of five notes of (1, 2, 3, 5, 6). There are dotted notes on both musical scales which represent moments of silence. Gamelan music uses a mode musical system called pathet. The system determines the characteristics of the composition based on the dominant notes including the arrangement of the order of the notes. The Gamelan composition is divided into various types, such as lancaran, ladrang, ketawang and others. Fig. 1 shows an example of a composition entitled Kembang Pete which is a type of lancaran composition with the lima musical mode and played on the pelog musical scale.

Parameters of the difficulty level of composition were determined by consulting with experts. Three experts were asked to define the parameters to be used in classifying compositions into basic, intermediate and advanced classes. Then, experts were asked to divide 50 compositions in the dataset into these three classes. Experts proposed the melody length and inter-notes distance to be used as parameters to determine the difficulty level of composition. The length of the melody is the number of beats in the composition. The greater number of beats increases the difficulty level in learning to play the instrument. Meanwhile, inter-notes distance is the distance between successive notes is the range between note values. For example, two successive notes of (1, 7) have a higher difficulty to play than (1, 3) because the distance of the first successive notes is 6 units and the distance of the second is 2 units.

Lancaran Kembang Pete, *Laras Pelog Pathet* Lima

6	5	3	2	1	2	3	5
6	5	3	2	1	2	3	5
6	5	6	1	6	5	3	2
5	6	5	4	2	1	6	5

Fig. 1. A Gamelan Music Composition Example.

The distance between two successive notes is the same, although the order is reversed. For example, (1, 7) and (7, 1) have the same distance of six units. So, the difficulty level of composition can be classified can be carried out using frequent pattern analysis rather than sequence analysis, including the addition of inter-notes distance analysis that appear in the note sequence. Moreover, the frequent pattern analysis was conducted using the Apriori algorithm, in which the melody length was also included in the calculation. The results of the frequent pattern analysis are then accumulated by the inter-notes distance value to classify compositions into basic, intermediate and advanced clasess.

The methodology used in this research consists of five stages, which are: data preparation, data representation, implementation of the Apriori algorithm, and classification where inter-notes distance analysis is performed.

A. Data Preparation

The dataset which consists of 50 Gamelan compositions in form of symbolic data were collected from a collection of music sheets. The music sheet data were then converted into text format. For example, the composition data shown in Fig. 1 is converted to (6, 5, 3, 2, 1, 2, 3, 5, 6, 5, 3, 2, 1, 2, 3, 5, 6, 5, 6, 1, 6, 5, 3, 2, 5, 6, 5, 4, 2, 1, 6, 5).

Some compositions contain dotted notes in their sequence, and the dotted notes were converted to 0 in order to support computational processing as in [22]. For example, as shown in Fig. 2, the composition entitled Balabak which is a type of ladrang composition with the lima musical mode and played on the pelog musical scale contains dotted notes. The composition is converted into a text format to (3, 2, 3, 1, 3, 2, 3, 5, 3, 2, 3, 1, 3, 2, 3, 5, 0, 0, 7, 6, 5, 4, 2, 1, 3, 2, 3, 1, 3, 2, 3, 5).

Ladrang Balabak, *Laras Pelog Pathet* Lima

3	2	3	1	3	2	3	5
3	2	3	1	3	2	3	5
•	•	7	6	5	4	2	1
3	2	3	1	3	2	3	5

Fig. 2. A Composition Containing Dotted Notes Example.

B. Data Representation

The dataset in this research uses symbolic format data from 50 compositions of ladrang style played in the pelog musical scale and the lima musical mode (the data set is included in the Appendix section, Table VIII). The music sheet data were represented by mapping the note sequence as the basis for determining transactions in frequent pattern analysis using the Apriori algorithm. Further, the results of the analysis were accumulated using inter-notes distance analysis. Frequent pattern analysis was performed per composition. Thus, each composition represents a set of transactions, and the note sequence is mapped as transactions within the composition, while notes of the musical scale are the transaction items. The note sequence mapping for the transactions was performed using the sliding window technique with the implementation on the k-itemset being performed using the following pseudocode:

D = a set of transaction containing note sequence data.

L = the length of D.

K = a K-itemset which is an itemset containing K successive notes.

T = transaction in D which represents a set of notes contained in the musical scale system.

```
for (z = 0; z < L; z++) {
        for (n = 0; n < K; n++) {
          T [z] [n] = D [z + n];
        }
}
```

The pseudocode above results in the last itemset having one less element length than the previous itemsets. In other words, all itemsets will have the same element length except the last itemset. Given a composition containing the note sequence of (3, 2, 3, 1, 3, 2, 3, 5, 3, 2, 3, 1, 3, 2, 3, 5, 0, 0, 7, 6, 5, 4, 2, 1, 3, 2, 3, 1, 3, 2, 3, 5) and the value of K is set to 2, then T will contain ((3, 2), (2, 3), (3, 1), (1, 3), (3, 2), …, (3, 2), (2, 3), (5)).

The melody has a repeating pattern; therefore, the last element in the last itemset will contain the first note by adding the following pseudocode:

```
n = L – K;
p = n + K;
while (n < p) {
        for (z = p – K; z < n; z++) {
          T [n] [K – (n – z)] = D [z - (p – K)]
        }
        n++
}
```

Continuing the previous example, then T will contain ((3, 2), (2, 3), (3, 1), (1, 3), (3, 2), …, (3, 2), (2, 3), (5, 3)), so all itemsets have the same length.

The inter-notes distance analysis was performed by subtracting the value of the higher note from the lower note, or subtracting the value of two same notes. For simplicity, the term unit is used as a measure of the value of the note and distance. There are seven notes and one dotted note denoted by the number 0 as a collection of items, which are (0, 1, 2, 3, 4, 5, 6, 7). Thus, the distance is measured with a scale from 0 to 7

units. Table I shows the measurement of the distance between successive notes.

TABLE I. TWO SUCCESSIVE NOTES DISTANCE MEASUREMENT

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0	1	2	3	4	5	6	7
1	1	0	1	2	3	4	5	6
2	2	1	0	1	2	3	4	5
3	3	2	1	0	1	2	3	4
4	4	3	2	1	0	1	2	3
5	5	4	3	2	1	0	1	2
6	6	5	4	3	2	1	0	1
7	7	6	5	4	3	2	1	0

Based on the description above, for example, given a K-itemset where K = 2 and let a transaction contains two successive notes of (3, 2), then the distance value will be 1 unit, and a transaction which contains two successive notes of (2, 3) also has the distance value of 1 unit. Meanwhile, a transaction of 2-itemset which contains two successive same notes, such as {0, 0}, or {4, 4}, has the distance value of 0 unit. The distance between successive notes is calculated by subtracting the elements of the first note by the elements of the second note. If any of the results are less than 0 or negative value, the result will be multiplied by -1 to make it a positive value. For example, two successive notes of (2, 5) will result in -3 from subtracting 2 – 5, and -3 will be multiplied by -1 to make a positive value of 3. For K-itemset where K is greater than 2, the distance between three or more successive notes is calculated by adding up all the distances between 2 successive notes on all elements in the itemset. The following is a pseudocode to calculate the inter-notes distance value where R represents the distance value between two successive notes, and S represents the total distance value of the itemset containing more than two successive notes.

```
n=0
while (n < L) {
        for (z = 0; z < K – 1; z++) {
          R [n] [z] = T [n] [z] – T [n] [z + 1];
          if (R [n] [z] < 0) {
                    R [n] [z] *= -1;
          }
          S [n] += R [n] [z];
        }
        n++;
}
```

Based on the formulas above, using an example of the note sequences of (3, 2, 3, 1, 3, 2, 3, 5, 3, 2, 3, 1, 3, 2, 3, 5, 0, 0, 7, 6, 5, 4, 2, 1, 3, 2, 3, 1, 3, 2, 3, 5), the data mapping into K-itemset of transactions, for example, where K = 2 and K = 3, and its distance value is as follows:

2-itemset

T = ((3, 2), (2, 3), (3, 1), (1, 3), (3, 2), …, (3, 2), (2, 3), (3, 5), (5, 3)).

S = ((1), (1), (2), (2), (1), …, (1), (1), (2), (2))

3-itemset

T = ((3, 2, 3), (2, 3, 1), (3, 1, 3), (1, 3, 2), (3, 2, 3), …, (3, 2, 3), (2, 3, 5), (3, 5, 3), (5, 3, 2)).

R = ((1, 1), (1, 2), (2, 2), (2, 1), (1, 1), …, (1, 1), (1, 2), (2, 2), (2, 1))

S = ((2), (3), (4), (3), (2), …, (2), (3), (4), (3))

C. Apriori Algorithm Implementation

The Apriori algorithm implementation was performed by mapping each composition as a separate set of transactions, and each transaction contains the same number of beats. In addition, the characteristics of the melody data structure, the classification problems encountered, and the data mapping carried out make the frequent itemset being applied to the 2-itemset pattern. Data mapping was performed based on two successive notes by applying the sliding window technique so as to produce patterns of {beat1, beat2}, {beat2, beat3}, {beat3, beat4}, ..., {last beat, beat1}. Thus, the data mapping can already represent the relationship between successive notes at each beat in the composition.

The following is an example of the Apriori algorithm implementation on a composition that is used as a dummy. Let D is the set of note sequence of the composition entitled Ladrang Balabak-Laras Pelog Pathet Slendro, where each transaction T in D contains items I which are elements of the set of notes in the pelog musical scale, and the transaction is mapped into two successive notes. Thus, with I, M, N, D, and T representing the set of items, the notes sequence, the number of beats in M, the set of transaction T containing note sequence data after data mapping, and the length of D, respectively, then:

I = (0, 1, 2, 3, 4, 5, 6, 7)

M = (3, 2, 3, 1, 3, 2, 3, 5, 3, 2, 3, 1, 3, 2, 3, 5, 0, 0, 7, 6, 5, 4, 2, 1, 3, 2, 3, 1, 3, 2, 3, 5)

N = 32

D = ((3, 2), (3, 1), (3, 2), (3, 5), (3, 2), (3, 1), (3, 2), (3, 5), (0, 0), (7, 6), (5, 4), (2, 1), (3, 2), (3, 1), (3, 2), (3, 5))

L = 16

The next data mapping uses a sliding window technique which doubles the number of beats and the length of the elements in D, as follows:

I = (0, 1, 2, 3, 4, 5, 6, 7)

M = (3, 2, 2, 3, 3, 1, 1, 3, 3, 2, …, 5, 3)

N = 64

D = ((3, 2), (2, 3), (3, 1), (1, 3), (3, 2), …, (3, 2), (2, 3), (3, 5), (5, 3)).

L = 32

TABLE II. TRANSACTION T DATA IN D

ID	0	1	2	3	4	5	6	7
1	0	0	1	1	0	0	0	0
2	0	0	1	1	0	0	0	0
3	0	1	0	1	0	0	0	0
4	0	1	0	1	0	0	0	0
5	0	0	1	1	0	0	0	0
6	0	0	1	1	0	0	0	0
7	0	0	0	1	0	1	0	0
…	…	…	…	…	…	…	…	…
31	0	0	0	1	0	1	0	0
32	0	0	0	1	0	1	0	0

Table II shows each transaction record in tabular data format with a sequential two-notes mapping in each transaction.

The next step is to calculate the frequent 2-itemset, including calculating the difficulty weight. The weight of itemset is measured based on the multiplication of the support value in each itemset with the distance between the notes in each itemset. The compositions in the dataset have varied melody lengths, and the longest is the composition with D containing 320 itemset, while the shortest containing 32 itemset.

The weight of itemset W result for each transaction in D is calculated by multiplying the support by the value of the distance between notes S, and dividing by the length of I (the set of items), which is W = (support x S) / the length of I, where I is the set of items containing eight notes. Next, the weight values in each transaction are summed to get the difficulty value. Table III shows the results of calculating the difficulty value of the composition, which is 0.2109.

The procedure performed in the example above was applied to all compositions in the dataset. Table IV shows an example of the results of calculating the difficulty value for the composition in the dataset.

TABLE III. 2-ITEMSET SUPPORT COUNT, TRANSACTION WEIGHT, AND THE DIFFICULTY VALUE OF THE COMPOSITION

Items	Count	Support	Dist.	Weight
{0, 0}	1	0.0312	0	0.0000
{2, 1}	1	0.0312	1	0.0039
{3, 1}	7	0.2187	2	0.0547
{3, 2}	12	0.375	1	0.0469
{4, 2}	1	0.0312	2	0.0078
{5, 0}	1	0.0312	5	0.0195
{5, 3}	5	0.1562	2	0.0391
{5, 4}	1	0.0312	1	0.0039
{6, 5}	1	0.0312	1	0.0039
{7, 0}	1	0.0312	7	0.0273
{7, 6}	1	0.0312	1	0.0039
Difficulty				0.2109

TABLE IV. THE DIFFICULTY VALUE OF COMPOSITIONS RESULTS EXAMPLES

ID	Length	Difficulty
1	80	0.2687
2	128	0.2695
3	32	0.2109
4	32	0.2109
5	96	0.2839
6	64	0.2734
7	64	0.3047
8	192	0.2331
9	128	0.2637
10	96	0.2682

D. Classification

Classification was performed using rules to discriminate compositions into three classes, which are basic, intermediate, and advanced. The difficulty level rules are defined by referring to the lowest value in the middle and advanced classes, the composition included in the basic class must be a composition that has a lower value than the lowest value found in the intermediate class, and the composition included in the intermediate class must be a composition that has a value between the lowest value found in the intermediate class to the lowest value found in the advanced class, while the composition included in the advanced class must be a composition that has a value greater than or equal to the lowest value found in that class. The following is the determination of the difficulty level rules:

IF value < minimum value in the intermediate class

THEN basic class

IF minimum value in the intermediate class ≥ value < minimum value in the advanced class

THEN intermediate class

IF value ≥ minimum value in the advanced class

THENadvanced class

Classification was performed by implementing the rules of the difficulty level based on the 50 compositions which have been divided into three classes by experts. Classification carried out by experts resulted in the division of 19, 17, and 14 compositions into basic, intermediate and advanced classes, respectively. Table V shows the results of the classification of compositions by experts and the value (difficulty level) of each composition. The data are displayed by sorting them based on the composition ID in ascending order.

There are different interpretations in the classification of the level of difficulty between the experts and the value of the level of difficulty. For example, the composition with ID 11 which is based on the assessment of the expert belongs to the basic class, has a difficulty value of 0.2969 that greater than the lowest difficulty value in the middle class, which is 0.2331 for composition ID 8. Moreover, all compositions in the intermediate class have a difficulty value greater than the lowest difficulty value in the advanced class. In this case, the

minimum and maximum values in each class need to be rearrangement. However, the rearrangement has an impact on the distribution of compositions by class. In other words, there will be a difference classification results between experts and the classifier.

The results of the classification by experts were used as a basis in determining the rules of the range of values for the level of difficulty for each class. Differences in the interpretation of the classification results by experts with the difficulty level value of each composition are resolved by finding the mean difficulty value in each class and inter-classes. Further, the mean difficulty values inter-classes are used as parameters to determine the value range of the difficulty level in each class. The following is data of the difficulty value of each composition based on its class collected from Table V sorted based on the value in ascending order. It is important to underline that the distribution of compositions into each class is determined by experts.

Basic = 0.1953, 0.1953, 0.1953, 0.2031, 0.2031, 0.2070, …, 0.3516)

Intermediate = (0.2331, 0.2484, 0.2500, 0.2559, 0.2578, …, 0.2734)

Advanced = 0.2406, 0.2578, 0.2695, 0.2760, 0.2773, …, 0.3984)

Next is calculating the mean value in each class. Mean difficulty value of the basic class is 0.2365, and the intermediate is 0.2614, while the advanced class is 0.2930. Finally, the minimum difficulty value in the intermediate class was determined based on the difficulty mean value calculation from the basic class mean value and the intermediate class mean value, which is (0.2365 + 0.2614) / 2 = 0.2489. Meanwhile, the minimum value of the difficulty level for the advanced class was determined based on the mean value calculation from the intermediate class mean value and the advanced class mean value, which is (0.2614 + 0. 2930) / 2 = 0.2772.

IF 0 < difficulty value < 0.2489

THEN basic class

IF 0.2489 ≤ difficulty value < 0.2772

THEN intermediate class

IF 0.2772 ≤ difficulty value ≤ 1

THEN advanced class

TABLE V. CLASSIFICATION RESULTS BY EXPERTS AND THE DIFFICULTY VALUES

Basic			**Intermediate**			**Advanced**		
ID	***Length***	***Value***	***ID***	***Length***	***Value***	***ID***	***Length***	***Value***
3	32	0.2109	1	80	0.2687	2	128	0.2695
4	32	0.2109	6	64	0.2734	5	96	0.2839
11	32	0.2969	8	192	0.2331	7	64	0.3047
14	64	0.1953	9	128	0.2637	13	96	0.2760
19	64	0.1953	10	96	0.2682	15	64	0.2813
20	64	0.2500	12	64	0.2578	24	160	0.3984
21	64	0.2422	16	64	0.2734	27	64	0.3281
22	96	0.2474	17	128	0.2559	30	96	0.2995
23	32	0.3516	18	64	0.2617	36	128	0.2773
26	64	0.2070	25	64	0.2734	38	320	0.2406
29	96	0.2083	28	128	0.2578	40	64	0.2969
32	64	0.2305	31	64	0.2656	41	128	0.2578
33	64	0.2383	35	64	0.2617	45	64	0.2813
34	32	0.3125	37	160	0.2484	46	96	0.3073
39	64	0.2500	42	64	0.2695			
43	64	0.2031	44	96	0.2500			
47	96	0.2031	50	96	0.2604			
48	128	0.2441						
49	96	0.1953						
Mean		**0.2365**	**Mean**		**0.2614**	**Mean**		**0.2930**

IV. Results and Discussion

The difficulty level rules above were evaluated by comparing the content of the three classes set by experts and the classifier. Compositions ID 11, 23 and 34 which are classified by the expert into the basic class are shifted out of the class. The compositions have difficulty value of 0.2969, 0.3516, and 0.3125, respectively, so they are shifted to the advanced class. Still in the basic class, both composition ID 20 and 39 have a difficulty value of 0.25, and they are shifted to the intermediate class. Meanwhile, in the advanced class, compositions ID 13 and 41 that have difficulty value of 0.2760 and 0.2578, respectively, are shifted to the intermediate class, while composition ID 38 that has the difficulty value of 0.2406 is shifted to the basic class.

The intermediate class showed positive results where there was only one classification difference between the experts and the classifier. Composition of ID 8 which is classified by experts into the intermediate class has a difficulty level value of 0.2331 that is in the basic class. Compositions classified by experts into basic classes have a number of beats less than or equal to128, while composition ID 8 is 192 beats long. Although the difficulty value of the composition indicates that it is in the basic class, the number of beats seems to be considered more by experts. This condition also applies to class shifts between basic and advanced classes.

Class shifts were found in several other cases so that the number of compositions in the basic, intermediate and advanced classes originally determined by the expert was 19, 17, and 14, changed to a total of 16, 21, 32 by the classifier, where changes also occur in some of its contents. Table VI shows the comparative results of the classification by experts and the classifier with B, I, A, E, and C, representing basic class, intermediate class, advanced class, experts, and the classifier, respectively.

Evaluation was carried out to measure the performance of the classifier by comparing the results of its classification to the classification by experts. The classifier performance was measured using a confusion matrix, as shown in Table VII. The results of the confusion matrix show that 14 of the 19 compositions classified by experts into basic classes can be identified by the classifier. Meanwhile, in the intermediate and advanced classes, 16 of the 17 compositions and 11 of the 14 compositions, respectively, can be identified by the classifier.

It is interesting to find out that three compositions of the basic class, which are compositions ID 11, 23 and 34, are shifted to the advanced class by the classifier. The three compositions have a length of 32 beats, and a melody with a length of 32 beats is the shortest composition in the dataset. The length of the melody seems to have a significant role compared to the inter-notes distance value. Another evidence can be seen in composition ID 39 which has a melody length of 320 beats. This composition is classified by experts into the advanced class. Meanwhile, it is shifted by the classifier to the basic class based on the difficulty value. The fact that all compositions classified by experts into the intermediate class have a melody length greater than 32 beats is another evidence.

Overall, the frequent and inter-notes distance analysis proposed in this research show good results in performing multi-class classifications for the difficulty level of composition based on the melodic pattern. The confusion matrix results were then calculated to measure accuracy, precision, and recall of the classifier performance. The performance of the classifier in the basic, intermediate and advanced classes achieved an accuracy of 87.5%, 80%, a precision of 73.7%, 94.1%, and 78.6%, and a recall of 87.5%, 80%, and 78.6%. Meanwhile, in total, the classifier's performance reached an accuracy, precision and recall value of 82%, 82.1%, and 82%.

The rule of difficulty classification is built on the frequency of notes and analysis of the distance between notes that is definitely found in all types of music, the difficulty classification model proposed in this study has a high potential to be applied to various types of music.

TABLE VI. Classification Results by the Classifier

ID	E	C	ID	E	C	ID	E	C	ID	E	C	ID	E	C
1	I	I	11	B	A	21	B	B	31	I	I	41	A	I
2	A	I	12	I	I	22	B	B	32	B	B	42	I	I
3	B	B	13	A	I	23	B	A	33	B	B	43	B	B
4	B	B	14	B	B	24	A	A	34	B	A	44	I	I
5	A	A	15	A	A	25	I	I	35	I	I	45	A	A
6	I	I	16	I	I	26	B	B	36	A	A	46	A	A
7	A	A	17	I	I	27	A	A	37	I	I	47	B	B
8	I	B	18	I	I	28	I	I	38	A	B	48	B	B
9	I	I	19	B	B	29	B	B	39	B	I	49	B	B
10	I	I	20	B	I	30	A	A	40	A	A	50	I	I

TABLE VII. CONFUSION MATRIX RESULTS

		Classifier		
		B	***I***	***A***
Experts	***B***	15	1	3
	I	1	16	0
	A	2	1	11

V. CONCLUSION AND FUTURE WORK

This research proposes a method for melodic patterns difficulty classification into basic, intermediate and advanced classes. The limited number of datasets is considered unsuitable for computing using a machine learning approach. Hence, instead of using a machine learning approach, classification was performed using a set of rules defined based on frequent pattern and inter-notes distance analysis. In successive notes, the forward notes and backward notes have the same pattern in the value of the distance between the notes. Thus, although the melody has a sequence or time series data type, the characteristics of the problem in the difficulty classification of melodic patterns are more suitable solved using frequent pattern analysis than sequence pattern. As a result, the proposed method is able to build a classifier to discriminate the difficulty level of composition with a good accuracy.

The future projection after completing this research is to implement it into adaptive learning to play musical instruments, and studies related to the classification of difficulty levels in composition are not yet popular even though this topic has relevance in adaptive learning. Fig. 3 shows a workflow diagram of future work where this research position is in the dashed-line box.

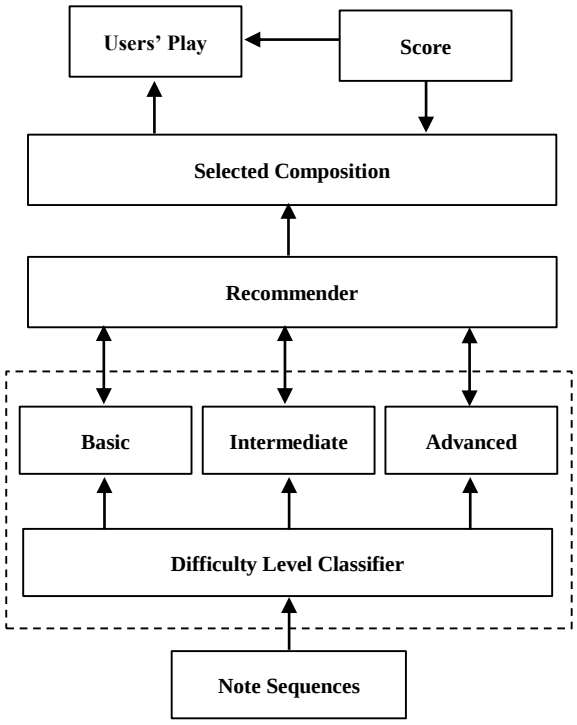


Fig. 3. Research Positions in Future Work.

ACKNOWLEDGMENT

This research is part of the research program of the Gamelan Metaverse lab developed by Universitas Dian Nuswantoro.

REFERENCES

[1] H. Khosravi, S. Sadiq, and D. Gasevi, "Development and Adoption of an Adaptive Learning System: Reflections and Lessons Learned," Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20). Association for Computing Machinery, New York, USA, 2020, pp. 58–64. Doi: 10.1145/3328778.3366900.

[2] V. Mirata, F. Hirt, P. Bergamin, and C. Van der Westhuizen, "Challenges and contexts in establishing adaptive learning in higher education: Findings from a Delphi study," International Journal of Educational Technology in Higher Education, vol. 17, no. 32, 2020, pp. 1-25. Doi: 10.1186/s41239-020-00209-y.

[3] A.V. Vega, O.C. Madrigal, and V. Kugurakova, "Approach of immersive adaptive learning for virtual reality simulator," Proceedings of 3rd Workshop on Advanced Virtual Environments and Education (WAVE2 2021), March 21-24, 2021, pp. 1-8. Doi: 10.5753/wave.2020.211635.

[4] W. Holmes, M. Bialik, and C. Fadel, "Artificial Intelligence in Education Promises and Implications for Teaching and Learning," The Center for Curriculum Redesign, Boston, MA, 2019.

[5] Y. Hao, K.S. Lee, S-T. Chen, and S.C. Sim, "An Evaluative Research of a Mobile Application for Middle School Students Struggling with English Vocabulary Learning," Computers in Human Behavior, 2019, vol. 95, pp. 208-216. Doi: 10.1016/j.chb.2018.10.013.

[6] M. Haug, P. Camps, T. Umland, and J-N. Voigt-Antons, "Assessing Differences in Flow State Induced by an Adaptive Music Learning Software," 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), 2020, pp. 1-4. Doi: 10.1109/QoMEX48832.2020.9123132.

[7] H. Peng, S. Ma, and J.M. Spector, "Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment," Smart Learning Environments, 2019, vol. 6, no. 9. Doi: 10.1186/s40561-019-0089-y.

[8] N. Yudana, and S. Wahyono, "The development of Gamelan learning media for android operating system," Indonesian journal of curriculum and educational technology studies, 2019, vol. 7, no. 2, pp. 64-71. Doi: 10.15294/ijcets.v7i2.29443.

[9] G. Presti, D. Adriano, F. Avanzini, A. Baratè, and L.A. Ludovico, "PhonHarp: A Hybrid Digital-Physical Musical Instrument for Mobile Phones Exploiting the Vocal Tract," Audio Mostly 2021 (AM '21). Association for Computing Machinery, New York, NY, USA, 2021, pp. 276-279. Doi: 10.1145/3478384.3478413.

[10] K. Lyu and R. Li, " Development of a Virtual Yangqin App with Unity Based on the Audio Object Pool Pattern," in: Shao X., Qian K., Zhou L., Wang X., Zhao Z. (eds), Proceedings of the 8th Conference on Sound and Music Technology. CSMT 2020. Lecture Notes in Electrical Engineering, vol 761. Springer, Singapore, 2021. Doi: 10.1007/978-981-16-1649-5_3.

[11] D. Lopes, G. Bernardes, L. Aly, and J. Forero, "Tumaracatu: An ubiquitous digital musical experience of maracatu," 11th Workshop on Ubiquitous Music (UbiMus 2021), Matosinhos, Portugal, 2021, October 12. Doi: 10.5281/zenodo.5564751.

[12] A.Z. Fanani, K. Hastuti, A.M. Syarif, and A.R. Mulyana, "Rule-based interactive learning application model on how to play music instruments," International Journal of Emerging Technologies in Learning, 2020, vol. 15, no. 15, pp. 52-63.

[13] S. Ruan, L. Jiang, J. Xu, B.J-K. Tham, Z. Qiu, Y. Zhu, E.L. Murnane, E. Brunskill, and J.A. Landay. "QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 2019, 1–13. Doi: 10.1145/3290605.3300587.

[14] M. Megahed and A. Mohammed, "Modeling adaptive E-Learning environment using facial expressions and fuzzy logic," Expert Systems

with Applications, 2020, vol. 157, no. 1. Doi: 10.1016/j.eswa.2020.113460.

[15] J. Yang, H. Huang, and S. Jin, "Mining Web Access Sequence with Improved Apriori Algorithm," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017, pp. 780-784. Doi: 10.1109/CSE-EUC.2017.154.

[16] C. Wang, and X. Zheng, "Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint," Evol. Intel., 2020, vol. 13, pp. 39-49. Doi: 10.1007/s12065-019-00234-5.

[17] S.B. Naik, and S. Khan, "Application of Association Rule Mining-Based Attribute Value Generation in Music Composition," in: V. Bhateja, S.C. Satapathy, C.M. Travieso-González, V.N.M. Aradhya (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, 2021, vol. 1407. Springer, Singapore. Doi: 10.1007/978-981-16-0171-2_36.

[18] K. Hastuti, A. Azhari, A. Musdholifah, and R. Supanggah, "Building Melodic Feature Knowledge of Gamelan Music Using Apriori Based on Functions in Sequence (AFiS) Algorithm," International Review on Computers and Software, 2016, vol. 11, no. 12, pp. 1127-1137. Doi: 10.15866/irecos.v11i12.10841.

[19] K.M. Jhang, M.C. Chang, T.Y. Lo, C.W. Lin, W.F. Wang, and H.H. Wu, "Using The Apriori Algorithm To Classify The Care Needs Of Patients With Different Types Of Dementia," Patient preference and adherence, 2019, vol. 13, pp. 1899–1912. Doi: 10.2147/PPA.S223816.

[20] X. Xie, G. Fu, Y. Xue, Z. Zhao, P. Chen, B. Lu, and S. Jiang, "Risk prediction and factors risk analysis based on IFOA-GRNN and apriori algorithms: Application of artificial intelligence in accident prevention," Process Safety and Environmental Protection, 2018, Doi: 10.1016/j.psep.2018.11.019.

[21] P. Rojanavasu, "Educational Data Analytics using Association Rule Mining and Classification," Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), 2019, pp. 142-145, Doi: 10.1109/ECTI-NCON.2019.8692274.

[22] A.M. Syarif, A. Azhari, S. Suprapto, and K. Hastuti, "Human and Computation-based Music Representation for Gamelan Music," Malaysian Journal of Music, 2020, vol. 9, pp. 82-100. Doi: 10.37134/mjm.vol9.7.2020.

APPENDIX

Dataset

TABLE VIII. COMPOSITIONS OF LADRANG STYLE OF THE PELOG MUSICAL SCALE AND THE LIMA MUSICAL MODE

No	Title	Note Sequence
1	Ladrang Arum Asih	0, 1, 0, 1, 2, 1, 2, 3, 0, 5, 1, 2, 3, 1, 2, 3, 0, 5, 1, 2, 3, 1, 2, 3, 1, 1, 0, 5, 6, 1, 2, 1, 0, 1, 0, 1, 2, 1, 2, 3, 0, 5, 1, 2, 3, 1, 2, 3, 1, 2, 3, 0, 5, 1, 2, 3, 1, 2, 3, 5, 5, 0, 6, 1, 6, 5, 4, 0, 4, 0, 4, 2, 4, 5, 6, 1, 6, 5, 4, 2, 4, 5, 4
2	Ladrang Babar Layar	6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 3, 2, 5, 3, 2, 5, 3, 2, 5, 3, 2, 5, 3, 2, 5, 2, 3, 5, 6, 5, 3, 6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 3, 2, 5, 3, 2, 5, 3, 2, 5, 3, 2, 5, 2, 3, 5, 6, 5, 4, 0, 4, 0, 4, 0, 4, 0, 1, 0, 1, 0, 1, 0, 1, 0, 5, 0, 1, 0, 5, 0, 1, 0, 5, 0, 4, 4, 6, 4, 5, 6, 1, 6, 5, 4, 6, 4, 5, 6, 1, 6, 5, 4, 6, 4, 5, 6, 1, 6, 5, 4, 6, 4, 5, 6, 1, 2, 3, 2, 1, 6, 5, 6, 3
3	Ladrang Balabak	3, 2, 3, 1, 3, 2, 3, 5, 3, 2, 3, 1, 3, 2, 3, 5, 0, 0, 7, 6, 5, 4, 2, 1, 3, 2, 3, 1, 3, 2, 3, 5
4	Ladrang Banten	3, 2, 3, 1, 3, 2, 3, 5, 6, 3, 6, 5, 3, 2, 3, 5, 6, 3, 6, 5, 3, 2, 3, 5, 6, 3, 6, 5, 3, 2, 3, 1
5	Ladrang Banyak Nglangi	0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 3, 3, 0, 0, 2, 1, 2, 3, 0, 1, 2, 3, 5, 1, 2, 3, 0, 1, 2, 3, 5, 1, 2, 3, 0, 1, 2, 3, 5, 1, 2, 3, 1, 1, 0, 0, 6, 5, 4, 5, 0, 2, 2, 0, 6, 5, 4, 5, 0, 2, 2, 0, 6, 5, 4, 5, 0, 2, 5, 4, 0, 2, 5, 4, 0, 2, 5, 4, 2, 1, 6, 5
6	Ladrang Bayemtur	0, 3, 5, 6, 3, 5, 3, 2, 0, 3, 5, 6, 3, 5, 3, 2, 0, 3, 3, 0, 3, 6, 3, 5, 3, 6, 3, 5, 3, 1, 3, 2, 0, 4, 4, 4, 2, 1, 2, 6, 0, 4, 4, 4, 2, 1, 2, 6, 0, 3, 3, 0, 3, 6, 3, 5, 3, 6, 3, 5, 3, 1, 3, 2
7	Ladrang Bedhati	0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 0, 0, 5, 6, 5, 3, 2, 1, 6, 5, 6, 1, 2, 3, 2, 1, 0, 6, 5, 6, 1, 2, 3, 1, 0, 6, 5, 6, 1, 2, 3, 1, 0, 6, 5, 6, 3, 2, 3, 1, 3, 2, 1, 6, 2, 1, 6, 5
8	Ladrang Bima Kurda	2, 1, 2, 1, 2, 5, 6, 1, 2, 1, 2, 1, 2, 5, 6, 1, 2, 1, 2, 1, 2, 5, 6, 1, 5, 6, 1, 2, 5, 3, 2, 1, 3, 3, 0, 0, 1, 2, 3, 2, 3, 1, 2, 0, 3, 2, 1, 2, 3, 1, 2, 0, 3, 2, 1, 2, 0, 1, 6, 0, 5, 6, 2, 1, 5, 5, 0, 0, 1, 6, 5, 3, 2, 3, 5, 0, 1, 6, 5, 1, 2, 3, 5, 0, 1, 6, 5, 3, 0, 2, 1, 0, 5, 6, 2, 1, 2, 1, 2, 1, 2, 5, 6, 1, 2, 1, 2, 1, 2, 5, 6, 1, 2, 1, 2, 5, 6, 1, 5, 6, 1, 2, 5, 3, 2, 1, 3, 3, 5, 3, 1, 2, 3, 2, 3, 1, 2, 3, 1, 2, 3, 2, 3, 1, 2, 3, 1, 2, 3, 2, 0, 1, 6, 0, 5, 6, 2, 1, 5, 5, 6, 5, 7, 6, 5, 3, 2, 3, 5, 6, 7, 6, 5, 3, 2, 3, 5, 6, 7, 6, 5, 3, 0, 2, 1, 0, 5, 6, 2, 1
9	Ladrang Blabak	0, 6, 5, 4, 2, 4, 6, 5, 1, 1, 0, 0, 2, 3, 2, 1, 5, 6, 1, 5, 6, 1, 2, 1, 3, 2, 1, 2, 0, 1, 2, 6, 0, 6, 5, 4, 2, 4, 6, 5, 7, 7, 0, 0, 7, 7, 6, 5, 2, 3, 5, 2, 3, 5, 6, 5, 7, 6, 5, 6, 5, 3, 2, 3, 6, 5, 2, 1, 6, 1, 3, 2, 7, 7, 0, 0, 7, 7, 6, 5, 2, 3, 5, 2, 3, 5, 6, 5, 7, 6, 5, 6, 5, 3, 2, 1, 6, 5, 2, 1, 6, 1, 3, 2, 5, 5, 0, 0, 5, 4, 2, 1, 5, 6, 1, 5, 6, 1, 2, 1, 3, 2, 1, 2, 0, 1, 2, 6
10	Ladrang Dhandanggulo Maskentar	0, 0, 6, 0, 6, 6, 1, 2, 0, 0, 2, 1, 3, 2, 1, 6, 0, 0, 6, 0, 5, 5, 6, 1, 3, 2, 1, 2, 0, 1, 6, 5, 0, 0, 0, 0, 1, 2, 1, 6, 2, 1, 5, 2, 0, 1, 0, 6, 3, 3, 0, 0, 6, 5, 3, 2, 0, 3, 2, 1, 6, 5, 3, 5, 2, 2, 0, 1, 3, 2, 1, 6, 2, 3, 2, 1, 6, 5, 3, 5, 2, 2, 0, 0, 5, 6, 5, 4, 6, 5, 2, 3, 2, 1, 2, 1
11	Ladrang Dhengklung	1, 6, 1, 2, 1, 6, 3, 5, 4, 2, 4, 5, 4, 2, 4, 1, 4, 2, 4, 5, 4, 2, 4, 1, 5, 6, 1, 2, 1, 6, 4, 5
12	Ladrang Dhudha Kondhang	3, 2, 5, 6, 1, 5, 6, 1, 3, 2, 5, 6, 1, 5, 6, 1, 3, 2, 5, 6, 1, 5, 6, 1, 2, 3, 5, 3, 2, 1, 2, 1, 3, 3, 0, 0, 3, 3, 5, 3, 5, 6, 7, 6, 5, 3, 2, 3, 5, 5, 0, 0, 7, 6, 5, 3, 1, 1, 0, 5, 6, 1, 2, 1
13	Ladrang Durma	0, 1, 1, 1, 6, 1, 2, 3, 0, 0, 3, 2, 0, 1, 6, 5, 1, 5, 0, 0, 5, 5, 0, 6, 1, 1, 0, 5, 6, 1, 2, 1, 5, 5, 0, 0, 5, 5, 3, 5, 6, 6, 7, 6, 5, 4, 2, 1, 5, 5, 0, 6, 5, 3, 1, 2, 3, 2, 1, 6, 5, 6, 1, 2, 0, 0, 2, 3, 5, 5, 3, 5, 6, 6, 5, 4, 2, 1, 6, 5, 1, 5, 0, 0, 5, 5, 0, 6, 1, 1, 0, 5, 6, 1, 2, 1
14	Ladrang Eling-Eling	6, 5, 3, 2, 1, 2, 3, 5, 6, 5, 3, 2, 1, 2, 3, 5, 1, 1, 0, 0, 1, 2, 3, 5, 3, 2, 3, 1, 3, 2, 6, 5, 6, 5, 2, 1, 3, 2, 6, 5, 6, 5, 2, 1, 3, 2, 6, 5, 2, 1, 2, 1, 3, 2, 6, 5, 3, 2, 3, 1, 3, 2, 6, 5

15	Ladrang Eling-Eling Subasiti	6, 5, 3, 2, 1, 2, 3, 5, 6, 5, 6, 1, 3, 2, 6, 5, 6, 5, 2, 1, 3, 2, 3, 1, 2, 1, 5, 6, 1, 1, 2, 1, 6, 1, 6, 1, 6, 1, 2, 1, 4, 5, 4, 5, 6, 1, 6, 5, 0, 3, 2, 1, 6, 5, 6, 1, 6, 6, 2, 1, 2, 6, 3, 5
16	Ladrang Glendheh	5, 6, 1, 0, 1, 3, 1, 2, 5, 6, 1, 0, 1, 3, 1, 2, 5, 6, 1, 0, 1, 3, 1, 2, 5, 6, 5, 4, 2, 1, 2, 1, 5, 5, 0, 0, 6, 4, 6, 5, 0, 5, 5, 5, 6, 4, 6, 5, 0, 0, 4, 0, 2, 0, 4, 0, 2, 4, 6, 5, 0, 4, 2, 1
17	Ladrang Golong	0, 6, 1, 2, 1, 6, 3, 5, 3, 2, 3, 0, 3, 6, 3, 5, 3, 2, 3, 0, 3, 6, 3, 5, 4, 2, 1, 2, 1, 6, 3, 5, 0, 0, 5, 2, 3, 5, 6, 5, 0, 0, 5, 6, 7, 7, 5, 6, 0, 6, 3, 5, 6, 7, 5, 6, 7, 5, 3, 2, 5, 6, 5, 3, 6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 6, 3, 6, 5, 3, 2, 3, 1, 6, 1, 2, 3, 5, 3, 6, 5, 3, 5, 3, 2, 3, 2, 3, 5, 6, 3, 5, 6, 7, 6, 7, 5, 6, 7, 6, 5, 3, 2, 5, 6, 5, 4, 2, 1, 2, 1, 3, 5, 3, 2, 1, 6, 3, 5
18	Ladrang Gudhasih	0, 6, 1, 2, 1, 6, 4, 5, 3, 1, 3, 2, 1, 6, 4, 5, 2, 2, 0, 0, 2, 3, 2, 1, 5, 6, 1, 2, 1, 6, 4, 5, 0, 5, 5, 5, 6, 4, 6, 5, 2, 4, 5, 6, 5, 4, 2, 1, 0, 2, 4, 5, 4, 2, 4, 1, 5, 6, 1, 2, 1, 6, 4, 5
19	Ladrang Gunung Kembar	0, 3, 2, 3, 5, 6, 5, 3, 0, 3, 2, 3, 5, 6, 5, 3, 0, 0, 3, 2, 3, 5, 6, 5, 0, 0, 3, 5, 3, 2, 3, 1, 0, 1, 2, 0, 2, 3, 2, 1, 0, 1, 2, 0, 2, 1, 3, 2, 0, 1, 6, 5, 6, 6, 5, 6, 3, 3, 2, 3, 5, 6, 5, 3
20	Ladrang Hastama	2, 1, 2, 4, 5, 4, 2, 1, 2, 1, 2, 4, 5, 4, 2, 1, 3, 2, 1, 2, 0, 1, 6, 5, 1, 5, 0, 6, 1, 0, 2, 1, 5, 5, 0, 0, 5, 5, 4, 5, 6, 6, 5, 6, 4, 5, 6, 5, 6, 5, 4, 2, 1, 6, 4, 5, 1, 5, 0, 6, 1, 0, 2, 1
21	Ladrang Jong Layar	0, 0, 0, 0, 2, 2, 3, 2, 0, 0, 3, 5, 0, 0, 3, 2, 0, 0, 3, 5, 0, 0, 3, 2, 5, 3, 2, 5, 2, 3, 5, 6, 0, 0, 0, 0, 6, 6, 5, 6, 0, 0, 3, 5, 0, 0, 3, 2, 0, 0, 3, 5, 0, 0, 3, 2, 5, 3, 2, 5, 2, 3, 5, 6
22	Ladrang Kagok	0, 1, 1, 1, 5, 6, 2, 1, 0, 1, 1, 1, 5, 6, 1, 2, 0, 0, 2, 4, 5, 0, 6, 5, 6, 6, 5, 4, 2, 1, 2, 1, 5, 5, 0, 0, 5, 5, 3, 5, 0, 0, 5, 6, 7, 6, 5, 6, 0, 6, 5, 3, 2, 2, 3, 2, 0, 0, 2, 4, 5, 0, 6, 5, 7, 6, 5, 6, 5, 4, 2, 1, 3, 2, 1, 2, 0, 1, 6, 5, 0, 6, 1, 2, 0, 1, 6, 5, 1, 1, 0, 5, 6, 1, 2, 1
23	Ladrang Kapirekta	0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 3, 5, 3, 5, 6, 1, 6, 5, 7, 6, 2, 4, 2, 1, 6, 5
24	Ladrang Kodhokan	0, 6, 0, 3, 0, 5, 0, 2, 0, 6, 0, 3, 0, 5, 0, 2, 0, 6, 0, 3, 0, 5, 0, 2, 0, 6, 0, 5, 0, 3, 0, 2, 0, 1, 3, 2, 5, 6, 1, 2, 0, 1, 3, 2, 5, 6, 1, 2, 0, 1, 3, 2, 5, 6, 1, 2, 0, 6, 0, 5, 0, 3, 0, 2, 0, 3, 3, 0, 3, 5, 2, 3, 5, 2, 3, 0, 3, 5, 2, 3, 5, 2, 3, 0, 3, 5, 2, 3, 0, 6, 0, 5, 0, 3, 0, 2, 0, 7, 0, 7, 0, 6, 0, 5, 0, 4, 0, 2, 0, 4, 0, 1, 0, 4, 0, 2, 0, 4, 0, 1, 0, 4, 0, 6, 0, 4, 0, 5, 0, 0, 5, 6, 7, 7, 6, 7, 0, 6, 5, 6, 7, 7, 6, 7, 0, 6, 5, 6, 7, 7, 6, 7, 0, 6, 0, 5, 0, 3, 0, 2
25	Ladrang Kombang Mara	5, 3, 2, 1, 5, 3, 2, 1, 5, 3, 2, 1, 2, 1, 6, 5, 6, 1, 6, 5, 6, 1, 6, 5, 6, 1, 2, 3, 5, 3, 2, 1, 5, 5, 0, 0, 4, 4, 2, 5, 0, 0, 1, 6, 5, 4, 2, 5, 0, 0, 1, 6, 5, 4, 2, 1, 6, 1, 2, 3, 5, 3, 2, 1
26	Ladrang Kudhawa	3, 2, 3, 1, 3, 2, 3, 1, 3, 2, 3, 1, 0, 2, 3, 5, 0, 0, 5, 6, 7, 7, 6, 5, 3, 2, 3, 5, 3, 2, 3, 1, 5, 5, 0, 0, 5, 5, 3, 5, 6, 5, 3, 2, 1, 2, 3, 5, 6, 5, 3, 2, 1, 6, 3, 5, 3, 2, 3, 5, 3, 2, 3, 1
27	Ladrang Kumara Maya	6, 1, 6, 2, 6, 1, 6, 5, 6, 1, 6, 2, 6, 1, 6, 5, 0, 5, 5, 5, 6, 4, 6, 5, 1, 2, 1, 6, 5, 4, 2, 1, 5, 6, 1, 6, 5, 4, 2, 1, 5, 6, 1, 6, 5, 4, 2, 1, 6, 6, 0, 0, 6, 5, 4, 2, 4, 5, 6, 5, 2, 1, 6, 5
28	Ladrang Langen Branta	0, 1, 0, 1, 6, 1, 2, 3, 5, 6, 5, 3, 2, 1, 2, 1, 0, 5, 5, 0, 5, 6, 1, 2, 3, 3, 5, 3, 2, 1, 2, 1, 0, 1, 0, 1, 6, 1, 2, 3, 5, 6, 5, 3, 2, 1, 2, 1, 0, 5, 5, 0, 5, 6, 1, 2, 3, 5, 3, 2, 1, 6, 3, 5, 6, 5, 6, 0, 6, 5, 2, 1, 3, 5, 3, 2, 1, 6, 3, 5, 0, 4, 4, 2, 4, 5, 2, 1, 3, 5, 3, 2, 1, 6, 3, 5, 6, 5, 6, 0, 6, 5, 2, 1, 3, 5, 3, 2, 1, 6, 3, 5, 0, 4, 4, 2, 4, 5, 2, 1, 3, 3, 5, 3, 2, 1, 2, 1
29	Ladrang Larastangis	0, 1, 1, 1, 2, 3, 2, 1, 0, 1, 1, 1, 2, 3, 2, 1, 0, 2, 1, 0, 2, 1, 6, 5, 0, 0, 5, 6, 1, 2, 3, 2, 0, 0, 0, 0, 2, 2, 3, 2, 0, 0, 2, 3, 2, 1, 2, 1, 0, 2, 1, 0, 2, 1, 6, 5, 0, 0, 5, 6, 1, 2, 3, 2, 5, 5, 0, 0, 5, 5, 3, 5, 0, 0, 5, 6, 7, 7, 6, 7, 0, 0, 0, 0, 7, 6, 5, 3, 0, 0, 2, 5, 0, 3, 2, 1
30	Ladrang Lebdajiwa	0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 1, 1, 0, 0, 5, 6, 1, 2, 1, 3, 1, 2, 0, 1, 6, 5, 0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 0, 0, 0, 0, 6, 4, 6, 5, 2, 4, 5, 6, 5, 4, 1, 2, 6, 6, 0, 0, 2, 1, 6, 5, 1, 2, 1, 6, 5, 4, 1, 2, 1, 1, 0, 0, 5, 6, 1, 2, 1, 3, 1, 2, 0, 1, 6, 5
31	Ladrang Manik Maninten	0, 1, 1, 1, 2, 3, 2, 1, 5, 5, 0, 0, 6, 1, 6, 5, 0, 5, 3, 5, 6, 1, 6, 5, 6, 1, 2, 1, 2, 3, 2, 1, 5, 5, 0, 0, 6, 1, 6, 5, 6, 1, 2, 1, 2, 3, 2, 1, 0, 1, 6, 1, 2, 3, 2, 1, 6, 6, 5, 4, 2, 1, 2, 1
32	Ladrang Maya	0, 1, 1, 1, 5, 6, 2, 1, 2, 1, 2, 3, 5, 3, 2, 1, 3, 2, 1, 2, 0, 1, 6, 5, 1, 5, 0, 6, 1, 0, 2, 1, 5, 5, 0, 0, 5, 5, 6, 5, 7, 6, 5, 6, 5, 4, 2, 1, 5, 5, 0, 6, 4, 5, 6, 5, 6, 6, 5, 4, 2, 1, 2, 1
33	Ladrang Menggak Layar	3, 2, 5, 6, 1, 5, 6, 1, 3, 2, 5, 6, 1, 5, 6, 1, 3, 2, 5, 6, 1, 5, 6, 1, 2, 3, 5, 3, 2, 1, 2, 1, 3, 3, 0, 0, 3, 3, 5, 3, 5, 6, 7, 6, 5, 3, 2, 3, 0, 3, 5, 6, 7, 6, 5, 3, 5, 6, 5, 3, 2, 1, 2, 1
34	Ladrang Nusantara	6, 5, 1, 6, 2, 1, 6, 5, 6, 5, 1, 6, 2, 1, 6, 5, 7, 6, 5, 6, 3, 5, 3, 2, 5, 3, 1, 6, 2, 1, 6, 5
35	Ladrang Obah	0, 6, 1, 2, 3, 1, 6, 5, 0, 1, 5, 6, 1, 2, 3, 2, 3, 2, 3, 5, 6, 5, 3, 2, 1, 3, 1, 2, 0, 1, 6, 5, 0, 5, 5, 5, 6, 4, 6, 5, 6, 5, 6, 1, 2, 1, 6, 5, 1, 2, 1, 6, 5, 4, 1, 2, 1, 3, 1, 2, 0, 1, 6, 5
36	Ladrang Pacarcina	0, 3, 2, 1, 6, 1, 3, 2, 0, 3, 2, 1, 6, 1, 2, 3, 0, 2, 5, 3, 0, 2, 5, 3, 5, 5, 6, 1, 2, 3, 1, 2, 5, 5, 0, 0, 5, 5, 3, 5, 0, 0, 5, 6, 7, 6, 5, 6, 0, 6, 5, 3, 2, 3, 6, 5, 7, 6, 5, 4, 2, 1, 2, 1, 0, 0, 0, 0, 1, 1, 2, 1, 3, 5, 3, 2, 0, 1, 6, 5, 0, 0, 2, 3, 5, 6, 7, 6, 3, 5, 6, 5, 3, 2, 1, 2, 6, 1, 6, 2, 6, 1, 6, 5, 6, 1, 6, 2, 6, 1, 6, 5, 6, 1, 6, 2, 6, 1, 6, 5, 3, 3, 6, 5, 3, 2, 1, 2

37	Ladrang Pasang Wetan	0, 0, 0, 0, 2, 2, 3, 2, 0, 0, 2, 3, 5, 6, 5, 3, 0, 0, 5, 3, 2, 1, 2, 6, 1, 2, 0, 6, 1, 2, 3, 2, 0, 0, 0, 0, 2, 2, 3, 2, 0, 0, 2, 3, 5, 6, 5, 3, 0, 0, 5, 3, 2, 1, 2, 6, 3, 5, 0, 2, 3, 5, 6, 5, 0, 0, 0, 0, 5, 5, 3, 5, 6, 6, 0, 3, 6, 5, 3, 5, 3, 2, 0, 3, 5, 6, 0, 3, 5, 6, 0, 3, 6, 5, 3, 2, 0, 0, 3, 0, 1, 2, 3, 2, 0, 2, 1, 6, 5, 6, 1, 6, 0, 0, 1, 6, 0, 0, 1, 6, 7, 7, 0, 0, 5, 6, 7, 6, 0, 2, 2, 0, 2, 3, 1, 2, 0, 0, 2, 3, 5, 6, 5, 3, 0, 0, 5, 3, 2, 1, 2, 6, 1, 2, 0, 6, 1, 2, 3, 2
38	Ladrang Playon	0, 6, 1, 2, 1, 6, 4, 5, 3, 3, 6, 5, 3, 2, 1, 6, 5, 6, 1, 2, 3, 2, 1, 2, 1, 6, 5, 4, 2, 4, 6, 5, 0, 5, 4, 2, 1, 2, 4, 5, 6, 5, 4, 2, 1, 2, 4, 5, 6, 5, 4, 2, 1, 2, 3, 2, 6, 6, 0, 7, 5, 6, 7, 6, 0, 6, 5, 4, 2, 2, 3, 2, 0, 0, 2, 4, 5, 0, 6, 5, 6, 5, 4, 2, 1, 0, 2, 1, 0, 2, 4, 5, 4, 2, 4, 1, 0, 2, 4, 5, 4, 2, 4, 1, 0, 2, 4, 5, 4, 2, 4, 1, 5, 5, 0, 0, 4, 5, 6, 5, 6, 5, 4, 2, 1, 2, 3, 2, 6, 6, 0, 7, 6, 5, 4, 5, 6, 5, 4, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 3, 3, 6, 5, 3, 2, 1, 6, 5, 6, 1, 2, 3, 2, 1, 2, 1, 6, 5, 4, 2, 4, 6, 5, 3, 3, 0, 0, 3, 3, 5, 3, 6, 5, 2, 1, 6, 1, 2, 3, 6, 5, 2, 1, 0, 5, 6, 1, 0, 4, 1, 2, 4, 5, 6, 5, 6, 5, 4, 2, 1, 6, 5, 6, 3, 5, 3, 2, 3, 2, 1, 6, 5, 6, 1, 2, 3, 2, 1, 2, 1, 6, 5, 4, 2, 4, 6, 5, 6, 5, 4, 2, 1, 1, 2, 1, 1, 6, 5, 4, 2, 4, 6, 5, 0, 2, 4, 5, 4, 2, 4, 1, 0, 2, 4, 5, 4, 2, 4, 1, 0, 2, 4, 5, 4, 2, 4, 1, 3, 3, 6, 5, 3, 2, 1, 6, 5, 6, 1, 2, 3, 2, 1, 2, 1, 6, 5, 4, 2, 4, 6, 5
39	Ladrang Prasaja	1, 5, 0, 6, 1, 1, 2, 1, 2, 3, 5, 3, 2, 1, 2, 1, 3, 3, 0, 0, 6, 5, 3, 2, 1, 3, 1, 2, 1, 6, 4, 5, 0, 5, 5, 5, 6, 4, 6, 5, 7, 6, 5, 6, 5, 4, 2, 1, 5, 5, 0, 6, 5, 3, 1, 2, 1, 3, 1, 2, 1, 6, 4, 5
40	Ladrang Pujiwidada	0, 6, 2, 1, 3, 2, 6, 5, 1, 6, 1, 5, 1, 6, 1, 2, 3, 6, 3, 5, 3, 1, 3, 2, 3, 1, 3, 2, 1, 6, 4, 5, 0, 5, 5, 5, 6, 4, 6, 5, 1, 2, 1, 6, 5, 4, 1, 2, 3, 6, 3, 5, 3, 1, 3, 2, 3, 1, 3, 2, 1, 6, 4, 5
41	Ladrang Randha Ngangsu	1, 2, 1, 6, 5, 6, 5, 0, 5, 6, 1, 2, 3, 2, 1, 0, 1, 2, 1, 6, 5, 6, 5, 0, 2, 4, 2, 4, 5, 6, 4, 5, 0, 6, 5, 4, 2, 4, 2, 0, 2, 4, 2, 4, 5, 6, 4, 5, 0, 4, 4, 5, 0, 4, 4, 5, 0, 1, 1, 2, 3, 2, 1, 0, 0, 0, 3, 2, 0, 1, 6, 5, 1, 5, 0, 6, 1, 0, 2, 1, 0, 1, 1, 1, 5, 6, 2, 1, 3, 2, 6, 5, 2, 4, 6, 5, 0, 6, 5, 6, 1, 1, 2, 1, 4, 4, 6, 5, 2, 4, 6, 5, 6, 6, 0, 0, 5, 6, 1, 6, 5, 2, 4, 5, 1, 1, 2, 1
42	Ladrang Raraskaton	2, 1, 2, 6, 2, 1, 6, 5, 2, 1, 2, 6, 2, 1, 6, 5, 2, 1, 2, 6, 2, 1, 6, 5, 2, 3, 5, 3, 3, 2, 1, 2, 2, 3, 0, 0, 6, 5, 3, 2, 6, 5, 3, 5, 3, 2, 1, 2, 3, 2, 1, 6, 2, 1, 6, 5, 2, 1, 2, 6, 2, 1, 6, 5
43	Ladrang Rasamulya	0, 1, 1, 1, 2, 3, 2, 1, 0, 1, 1, 1, 2, 3, 2, 1, 5, 6, 1, 0, 2, 1, 6, 5, 0, 0, 5, 6, 1, 1, 2, 1, 5, 5, 0, 0, 5, 5, 6, 5, 7, 6, 5, 6, 5, 4, 1, 2, 0, 0, 2, 4, 5, 0, 6, 5, 6, 6, 5, 4, 2, 1, 2, 1
44	Ladrang Retna Kedhiri	0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 3, 1, 3, 2, 3, 1, 3, 2, 5, 6, 1, 2, 1, 6, 4, 5, 6, 5, 4, 2, 1, 2, 4, 5, 6, 5, 4, 2, 1, 2, 4, 5, 6, 5, 4, 2, 1, 2, 3, 2, 6, 6, 0, 7, 6, 5, 4, 5, 7, 6, 5, 6, 5, 4, 2, 1, 0, 2, 4, 5, 4, 2, 4, 1, 0, 2, 4, 5, 4, 2, 4, 1, 5, 6, 1, 2, 1, 6, 4, 5
45	Ladrang Retnaningsih	0, 6, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 4, 5, 1, 1, 0, 0, 5, 6, 1, 2, 1, 3, 1, 2, 0, 1, 6, 5, 0, 5, 5, 5, 6, 4, 6, 5, 0, 5, 5, 5, 6, 4, 6, 5, 6, 5, 4, 2, 1, 1, 2, 1, 5, 6, 1, 2, 0, 1, 6, 5
46	Ladrang Santi Mulya	6, 1, 6, 5, 6, 1, 6, 5, 2, 4, 5, 6, 5, 4, 2, 1, 6, 5, 6, 1, 6, 5, 6, 1, 2, 3, 2, 1, 2, 1, 6, 5, 2, 1, 6, 5, 2, 1, 6, 5, 0, 6, 3, 2, 1, 6, 3, 5, 0, 0, 5, 0, 5, 3, 2, 1, 2, 6, 2, 1, 3, 2, 6, 5, 6, 6, 0, 0, 4, 5, 6, 1, 2, 1, 6, 5, 4, 5, 6, 1, 3, 2, 1, 2, 1, 6, 4, 5, 0, 6, 1, 2, 1, 6, 3, 5
47	Ladrang Sarilaya	2, 2, 0, 0, 2, 2, 3, 2, 0, 0, 2, 3, 5, 6, 5, 3, 0, 3, 2, 1, 3, 2, 1, 2, 3, 3, 2, 1, 6, 5, 3, 5, 1, 1, 0, 0, 1, 1, 2, 1, 3, 2, 1, 2, 0, 1, 6, 5, 0, 0, 2, 3, 5, 6, 5, 6, 3, 5, 6, 5, 3, 2, 1, 2, 6, 6, 0, 0, 6, 6, 5, 6, 0, 0, 7, 6, 5, 3, 2, 3, 0, 3, 2, 1, 3, 2, 1, 2, 3, 3, 2, 1, 6, 5, 3, 5
48	Ladrang Sayang Gunung	0, 6, 1, 2, 0, 1, 6, 5, 0, 6, 1, 2, 0, 1, 6, 5, 1, 1, 0, 0, 5, 6, 1, 2, 3, 3, 1, 2, 0, 1, 6, 5, 3, 3, 0, 0, 2, 1, 2, 3, 6, 5, 3, 2, 3, 1, 2, 3, 6, 5, 3, 2, 3, 1, 2, 3, 1, 1, 3, 2, 0, 1, 6, 5, 3, 3, 0, 0, 2, 1, 2, 3, 6, 5, 3, 2, 3, 1, 2, 3, 1, 2, 3, 5, 0, 3, 2, 1, 3, 2, 1, 2, 0, 1, 6, 5, 1, 2, 1, 6, 5, 4, 2, 1, 3, 2, 1, 2, 0, 1, 2, 6, 0, 0, 6, 5, 6, 1, 2, 1, 0, 3, 0, 2, 0, 1, 6, 5
49	Ladrang Sembawa	0, 1, 1, 1, 2, 3, 2, 1, 0, 1, 1, 1, 2, 3, 5, 3, 0, 3, 5, 6, 7, 6, 5, 3, 5, 3, 2, 3, 2, 1, 2, 1, 5, 5, 0, 0, 5, 5, 3, 5, 0, 0, 5, 6, 7, 6, 5, 6, 0, 6, 5, 3, 6, 5, 3, 5, 7, 6, 2, 1, 2, 3, 5, 3, 0, 3, 3, 3, 2, 1, 2, 1, 0, 1, 1, 1, 2, 3, 5, 3, 0, 3, 5, 6, 7, 6, 5, 3, 5, 3, 2, 3, 2, 1, 2, 1
50	Ladrang Singa-Singa	0, 6, 5, 6, 1, 2, 1, 6, 0, 6, 5, 6, 1, 2, 1, 6, 0, 6, 5, 6, 1, 2, 1, 6, 5, 6, 5, 4, 2, 4, 5, 6, 0, 6, 6, 6, 5, 4, 2, 1, 0, 1, 2, 4, 5, 4, 2, 1, 0, 1, 2, 4, 5, 4, 2, 1, 3, 2, 1, 6, 2, 4, 2, 1, 0, 0, 1, 2, 3, 2, 1, 2, 0, 2, 1, 0, 1, 2, 1, 6, 0, 6, 5, 6, 1, 2, 1, 6, 5, 6, 5, 4, 2, 4, 5, 6

Machine Learning: Assisted Cardiovascular Diseases Diagnosis

Aseel Alfaidi[1], Reem Aljuhani[2], Bushra Alshehri[3], Hajer Alwadei[4], Sahar Sabbeh[5]
Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, KSA[1, 2, 3, 4, 5]
Faculty of Computer Sciences and Artificial Intelligence, Benha University, Egypt[5]

***Abstract*—Detecting cardiovascular problems during their early stages is one of the great difficulties facing physicians. Cardiovascular diseases contribute to the deaths of around 18 million patients every year worldwide. That's why heart disease is a critical worry that must be addressed. However, it can be difficult to detect heart disease because of the multiple factors that affect health, such as high blood pressure, increased cholesterol, abnormal pulse rate, and many other factors. Therefore, the field of artificial intelligence can be instrumental in detecting diseases early on and finding an appropriate solution. This paper proposes a model for diagnosing the probability of an individual having cardiovascular illness by employing Machine Learning (ML) models. The experiments were executed using seven algorithms, and a public dataset of cardiovascular disease was used to train the models. A Chi-square test was used to identify the most important features to predict cardiovascular disease. The experiment results showed that Multi-Layer Perceptron gives the highest accuracy of disease prediction at 87.23%.**

***Keywords*—Cardiovascular diseases; artificial intelligence; prediction; multi-layer perceptron**

I. Introduction

The human heart is the most critical component in the body, whose main task is to pump blood to all body parts. The heart is at the center of the circulatory system and is a network of blood vessels such as arteries, veins, and capillaries [1]. Any component in the body is exposed to diseases and injuries, but the heart is among the body's major organs. As its damage may threaten human life, and its diseases or injuries cannot be easily overlooked.

Heart disease disrupts the heart's regular electrical system and pumping functions. Shortness of breath, physical weakness, swollen feet, and weariness can indicate heart disease [2]. Causes threatening human heart health include high cholesterol, smoking, lack of physical activity, and increased blood pressure [3].

Cardiovascular diseases are a group of disorders brought on by cardiac issues. According to the World Health Organization [4], the leading reason for death is the cardiovascular disease as it causes the death of 18 million patients each year, around 32% of the deaths around the world. Hence, cardiovascular diseases are viewed as a significant health concern. There are several types of cardiac illness, the most prevalent are heart.

Angiography is the method that most doctors use to diagnose cardiovascular patients. However, this diagnosing process requires analyzing many factors, which is also considered an expensive procedure, especially, in developing countries that suffer from a scarcity of diagnostic devices, doctors, and other resources [5][6].

As the number of deaths caused by cardiovascular diseases rises every day, the prediction of these diseases has become one of the most crucial subjects in the medical field. Prediction helps to detect disease in its early stages, thus, reduce the risk of sickness, or treat it most effectively.

Machine learning (ML) techniques in the domain of medical diagnosis are continuously expanding. This can be attributed mostly to advancements in disease classification and recognition, which can provide data that support medical specialists in the early discovery and diagnosis of diseases, thus maintaining human health and reducing the death rate. The classification algorithms are ML learning approaches that are often used to identify the probability of disease occurrence [7] [8]. Therefore, this paper aims to build a classification model to predict cardiovascular disease using real world dataset of cardiovascular patients.

The focus of ML is to develop systems that can make predictions based on experience [8].There are three types of machine learning techniques. First, the supervised learning, where the model is trained using labeled data, and the performance of the model is evaluated using test data. Supervised learning usually includes classification and regression problems. The second type is unsupervised learning, in which the data is not labelled and the model tries to discover the hidden patterns that may exist in the data. It derives conclusions from datasets to characterize hidden knowledge after exploring data. A clustering approach is an example of unsupervised learning [9]. The third type is reinforcement learning, which neither makes use of labelled data, nor the findings are related to the data. It is concerned with how intelligent agents can take actions in an environment [10].

The classification algorithms are ML approaches that are often used to identify the probability of disease occurrence[7][8]. It is a prominent machine learning technique that uses a model inferred from training data to predict the class of new samples [11][12]. Also, classification is a supervised learning concept that categorizes a set of data into classes [12]. This paper aims to build a classification model to predict cardiovascular disease using real world dataset of cardiovascular patients.

In this paper, we applied seven different classification algorithms to predict cardiovascular disease and determine the

best algorithm among them, namely, logistic regression (LR), decision tree, random forest (RF), naïve Bayesian (NB), k-nearest neighbor (KNN), support vector machine (SVM), and multi-layer perceptron (MLP).

The rest of this paper is organized as follows: Section 2 presents the related works in this filed. Section 3 describes in detail the research methodology, datasets, data preprocessing, and data analysis. Section 4 presents and discusses the results. Section 5 concludes the paper and provides the scope of future work.

II. Related Work

Researchers have suggested several possible ways to predict heart disease using various machine learning algorithms. The Cleveland Heart Disease dataset is the most common dataset used in heart disease prediction papers that are presented in this literature.

Rani et al. [13] proposed a decision system utilizing machine learning for cardio disease prediction based on a patient's clinical parameters. Their results indicated that the RF model had the best accuracy at 86.60%. Motarwar et al. [14] proposed a framework to predict the possibility of heart disease using various algorithms. They also found that RF achieved the best accuracy at 95.08%. Shah et al. [7] used several methods, such as the DT, NB, KNN, and RF algorithms. The results showed that the KNN algorithm had the greatest accuracy score at 90.7%.

Vijayashreea et al. [15] suggested a new fitness function for particle swarm optimization (PSO) using SVM. They created a new function based on identifying optimal weight population diversity and tuning for determining optimal weights. The SVM classifier's high performance was demonstrated using Receiver Operating Characteristic (ROC) analysis. In addition, they demonstrated the application of the suggested PSO-SVM–based feature selection technique for predicting heart disease. In addition, the SVM classifier was compared to other well-known classifier methods, such as NB, RF, and MLP.

Atallah and Al-Mousa [16] suggested using the complex voting ensemble method, and the outcome of the predictions is determined by a majority vote among all models. Consequently, the model attained 90% accuracy, which successfully exceeded the accuracy of each classifier. Besides the Cleveland Heart Disease dataset, Rao et al. [17] used the Switzerland, Hungarian, and Long Beach datasets. Using different algorithms for each dataset, the results showed that the highest accuracies were 86.81%, 98.30%, 84.26%, and 82.20% for Cleveland, Switzerland, Hungarian, and Long Beach, respectively.

Mohan et al. [18] developed a method that aims to improve the accuracy of cardiovascular disease prediction by applying machine learning techniques to find essential features. Feature selection depended on the machine learning techniques used, which included NB, generalized linear models, linear regression, deep learning, DT, RF, and SVM. The proposed hybrid random forest and linear model method was shown to be very accurate at predicting heart disease, with an accuracy of 88.7%. Another hybrid predictive system was proposed by Haq et al. [19] to diagnose heart disease. The authors used seven popular machine learning algorithms: LR, KNN, ANN, SVM, NB, DT, and RF. As a result, they concluded that LR had the best accuracy for predicting heart disease, with an accuracy of 89%. Kavitha et al. [20] proposed a heart disease prediction model using a hybrid approach. They implemented the model using three machine learning algorithms: RF, DT, and a hybrid of the two. Results showed a highest accuracy of 88.7%, achieved by the hybrid model.

Lakshmanarao et al. [21] suggested heart disease prediction using an ensemble classifier model. Two datasets were used in their study. First, they applied two feature selection techniques, namely Analysis of Variance (ANOVA) for F-value and mutual information. Based on these two techniques, they determined the best features. Nowshad et al. [22] gathered data in Bangladesh's Sylhet district by visiting hospitals and healthcare businesses in person to create a good questionnaire for heart disease prediction. There are 564 instances and 18 attributes in their dataset. The SVM produced the best results, with an accuracy level of 91%.

In contrast to previous studies, we used the cardiovascular disease dataset. To the best of our knowledge, this is the first study using this dataset which includes 70000 patients and 11 features. Also, we have applied different ML algorithms to determine the best for obtaining precise results for predicting cardiovascular disease.

III. Research Methodology

This section presents the methodology followed by researchers for the prediction of cardiovascular disease using machine learning algorithms. Fig. 1 illustrates our methodology process flow. First data are pre-processed, most informative features are selected, the resulting data are fed into different classification models and finally, performance is evaluated.

A. Dataset

The Cardiovascular Disease dataset obtained from the Kaggle repository [23] was used. The dataset has a sample size of 70000 patients and 11 features. Table I displays the feature dataset details and an explanation of each feature.

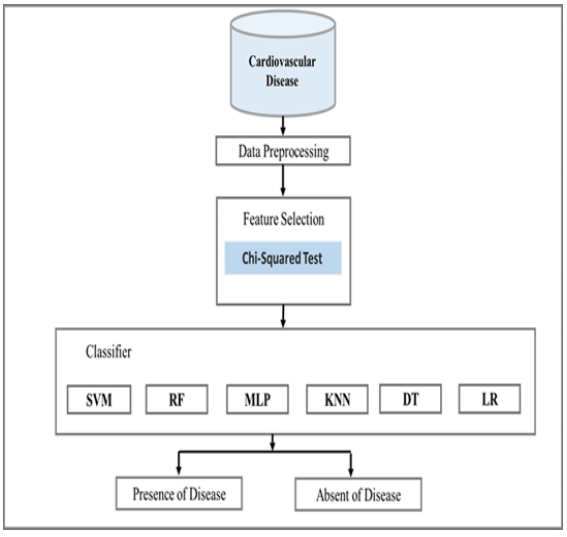


Fig. 1. Methodology Proposed to Predict Cardiovascular Disease.

TABLE I. INFORMATION IN THE DATASET

Feature name	Type	Description
Age	Discrete	Number of days
Gender	Discrete	Female: 2, Male: 1
Height	Continuous	In cm, Max = 250, Min = 55
Weight	Continuous	In kg, Max = 200, Min = 10
Systolic blood pressure	Discrete	Max = 16020, Min = −150
Diastolic blood pressure	Discrete	Max =11000, Min = −70
Cholesterol	Discrete	1: normal, 2: above normal, 3: well above normal
Glucose	Discrete	1: normal, 2: above normal, 3: well above normal
Smoking	Discrete	present: 1, absent: 0
Alcohol intake	Discrete	present: 1, absent: 0
Physical activity	Discrete	present: 1, absent: 0
Cardiovascular disease	Discrete	present: 1, absent: 0

B. Data Pre-processing

The Pre-processing of the dataset for a machine learning model is necessary for efficiency. We suggest in this study the pre-processing techniques of removing anomalies (outliers) and applying standard scaler to the dataset for showing the models efficiency and obtaining an acceptable and reliable accuracy for predicting the disease. After that, we used 68733 records from the dataset. We also modified some features of the dataset to best identify the factors that most influence cardiovascular disease as follows:

- Weight and height were merged into one feature in Body Mass Index (BMI): calculates body fat percentage based on height and weight.
- Features of the dataset were transformed while maintaining the information to make it more comprehensible. In this dataset, age was converted from days to years, and the gender feature was converted to binary.
- Values out of range (outliers) were removed in the.

C. Explanatory Data Analytics

The explanatory data analytics aims to use statistical and/or visual techniques to get insights into data sparsity, correlation, distribution, …etc.

The pie chart in Fig. 2A displays the gender distribution of the dataset, male 65.13% and female 34.87%. Fig. 2B shows the relationship of the gender feature to disease, indicating that the average number of females with cardiovascular disease is more than that of males.

Fig. 3 shows features relationships with the target feature. Fig. 3A indicates the cholesterol feature. Cholesterol is a waxy substance found in the blood, and high blood cholesterol is one of the heart diseases factors [24]. In the dataset, the rates of infection range between 1) normal and 3) well above normal; 30 well above normal has the highest occurrence of disease.

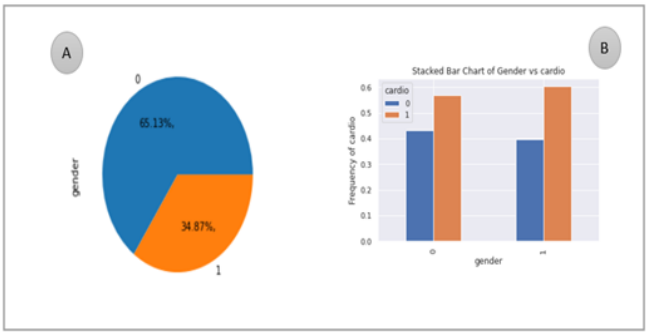


Fig. 2. Gender Feature Distribution of the Dataset.

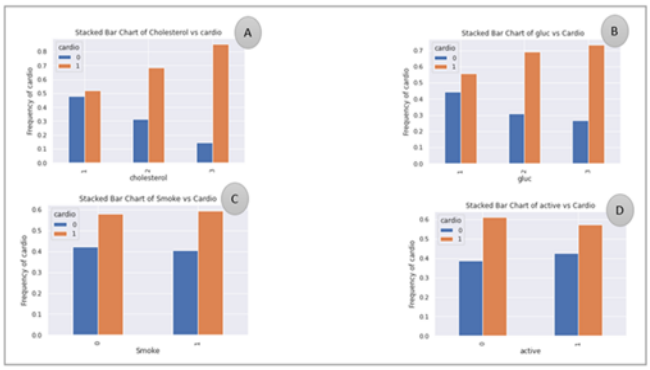


Fig. 3. Feature Distribution.

Fig. 3B shows the relationship between glucose and cardiovascular disease. High glucose is a known factor in cardiovascular disease [24]. The rates of glucose range between 1: normal and 3: well above normal in the dataset. Cardiovascular disease is most highly represented in 3: well above normal.

Indicated in Fig. 3C is the effect of smoking on the heart, a strongly linked known factor that causes cardiovascular disease [24]. The dataset indicates that it may have an effect in some cases. Finally, in Fig. 3D, it is shown that physical activity has a little effect on cardiovascular disease.

As demonstrated in Fig. 4, data visualization is utilized for discrete features to preview the distribution in the data. As shown in Fig. 4A, age data is distributed between 35 and 65 years. For the systolic blood pressure (ap_hi) attribute in Fig. 3B, we note a distribution from less than 100 to 200. Systolic blood pressure represents the heart's force on artery walls each time it beats [25]. We present the diastolic blood pressure (ap_lo) attribute in Fig. 4C, distributed between 50 and 125. Diastolic blood pressure measures the pressure on the walls of arteries between heartbeats [26].

Between each feature and the target feature(cardio), a correlation value was determined as shown in Table II. We note that the features that are positively correlated with the target feature (cardio) are ap_hi, ap_lo, age, cholesterol, BMI, gluc, gender, alco, and smoke; the active feature is the negatively correlated feature with the target feature (cardio).

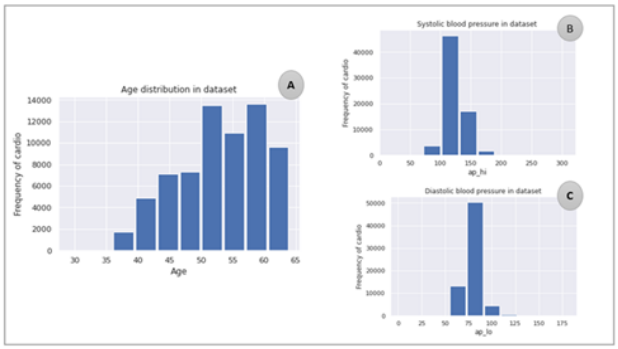


Fig. 4. Discrete Feature Distribution.

TABLE II. FEATURE AND CORRELATION VALUE OF THE DATASET

Feature code	Correlation value
Ap_hi	0.625865
Ap_lo	0.541238
Age	0.268197
Cholesterol	0.229901
BMI	0.217294
Gluc	0.111875
Gender	0.033391
Alco	0.018351
Smoke	0.008398
Id	0.004876
Active	-0.031358

D. Feature Selection

Feature selection is an essential step before data is fed to the classification model. It aims to reduce dimensionality by selecting the most informative feature that can contribute positively to the performance of the models.

In this study, for feature selection, we applied the filter method that uses variable ranking approaches as the primary criterion for ordering variables [27]. It is also a popular feature selection method in machine learning techniques and has achieved success for practical applications [27]. Moreover, we selected from the Chi-Squared test method. It uses statistical techniques to evaluate the relationship between the features and the target variables. In addition, it is used when a feature is tested and the target variable in the classification problem [27].

E. Machine Learning Models

For our experiment, we applied seven machine learning techniques namely, logistic regression, decision tree, random forest, naïve Bayesian, k-nearest neighbors, support vector machine, and multi-layer perceptron.

1) Logistic Regression (LR): a predictive analysis technique based on the concept of probability. LR can be compared to the Linear Regression model, LR, on the other hand, employs a more complicated computation: the sigmoid function, commonly known as a logistic function. In the context of LR, the cost function's range is from zero to one. Linear functions cannot be employed since their values can be less than zero or larger than one [22].

2) Random Forest (RF): The RF algorithm is a supervised classification approach. In RF, a forest is formed by several trees, each of which emits a class expectation, with the class with the most votes becoming the model's forecast. The bigger the number of trees in the RF classifier, the more accurate it is. It can be used for a variety of tasks, including classification and regression, but it shines when it comes to classification and dealing with missing information [28].

3) Decision Tree (DT): This is a classification algorithm that can be used to classify both category and numerical data. It has a type of structure that resembles a tree. DT is a primary and commonly used method for dealing with medical data. The data in a tree-shaped graph is simple to build and analyze. We used the DT classification method since it is one of the best and most widely used controlled learning strategies[29]. It is simple to build a stable decision tree for a given data collection.

4) Naïve Bayes (NB): We also use the NB classifier, a machine learning approach for identifying and forecasting the probability of an occurrence, is also used. Every NB classifier presupposes that a features value is different from the values of any other features in the class variables [22].

5) K-Nearest Neighbour (KNN): That is one of the most fundamental regression and classification machine learning approaches. The data is employed in KNN calculations, which use resemblance measures to characterize new points (e.g., distance function). In a nutshell, the KNN computation assumes the closeness of the comparable objects. In KNN, classification is made by taking into account the majority vote of its neighbours. The data point is labelled with the class that has most of its neighbours [30]. The selection of k and the accuracy may increase as the number of nearest neighbours increases.

6) Support Vector Machine (SVM): That can be used to perform regression and classification tasks. To implement SVM, we first represent each data item in our model as a number of features, with each component's estimation corresponding to a specific coordinate. The data is then classified by determining hyper-plane, which best separates the classes [7].

7) Multi-Layer Perceptron (MLP): Recently, it has been demonstrated that neural networks, specifically MLP, are excellent alternatives to more traditional statistical methodologies. It has been demonstrated that MLP may be trained to resemble almost any smooth, measurable function [31]. Unlike other statistical techniques, MLP makes no assumptions about the distribution of data. When given new, unknown inputs, it can represent extremely nonlinear functions and be trained to generalize appropriately. These characteristics make the MLP an appealing option for constructing numerical models as well as choosing among statistical approaches [31].

F. Evaluation Metrics

We suggest multiple ways to evaluate the classifiers as shown in Table III, to determine the appropriate model to predict disease.

TABLE III. EVALUATION METRICS

Definition	Equation
Accuracy: It is an assessment of a system ability to make correct predictions.	Accuracy = $(\frac{\text{Correct predictions}}{\text{Total predictions}}) \times 100$
Sensitivity: It is an assessment measures the ability of a system to predict positive outcomes correctly.	Sensitivity = $(\frac{\text{True positives}}{\text{True positives + false negatives}}) \times 100$.
Specificity : It is an assessment measures the ability of a system to predict negative outcomes correctly.	Specificity = $(\frac{\text{True negatives}}{\text{True negatives + false positives}}) \times 100$.
Precision: It is an assessment measures of a system to to the relevant results.	Precision = $(\frac{\text{True positives}}{\text{True positives + false positives}}) \times 100$.
F-measure: Is the sum of the results of measuring accuracy and sensitivity	F-measure = $2 \times (\frac{\text{Sensitivity} \times \text{precision}}{\text{Sensitivity + precision}})$

IV. RESULTS AND DISCUSSION

For our experiments, data were divided into training and testing sets with proportions of 70% and 30%, respectively. The classification was performed using the medical biomarkers available in the dataset, and class 1 means that the individual has a disease, while class 0 means that the person is disease-free.

Our first experiment targets the evaluating of the optimal number of features that can achieve the best accuracy among all models. To find that number, we assessed the accuracy obtained with each subset from one to ten combinations of features. Applying the Chi-squared method, the selection of features will be based on their rank that is determined by their scores. Tables IV and V shows all the combination of features and their corresponding accuracy by each classifier. As a result, with five selected features the highest accuracy was obtained by both MLP, LR, SVM, and RF with 87.2%, 85.5% 86.6%, and 86.0% respectively, whereas with all selected the highest accuracy was obtained by MLP with 87.2%.

TABLE IV. PERFORMANCE OF CLASSIFIERS ON DIFFERENT NUMBERS OF FEATURES

Number of features selected	Accuracy obtained by each model (%)						
	MLP	LR	SVM	RF	KNN	NB	DT
5	87.2	85.5	86.6	86	84.6	83.4	85.9
7	87.1	85.4	86.5	85.6	84.7	83.4	85.4
9	87.1	85.4	86.5	85.5	84.7	83.4	85.4

TABLE V. THE NUMBER OF FEATURES WITH CORRESPONDING SCORE

Number of features	Name of feature	Score		Gender	49.914749
1	Ap_hi	59514.160186		Alco	21.911426
2	Ap_hi	59514.160186	**9**	Ap_hi	59514.160186
	Ap_lo	23057.693730		Ap_lo	23057.693730
3	Ap_hi	59514.160186		Bmi	4436.047195
	Ap_lo	23057.693730		Age	4289.087958
	Bmi	4436.047195		Cholesterol	1226.927745
4	Ap_hi	59514.160186		Gluc	229.196698
	Ap_lo	23057.693730		Gender	49.914749
	Bmi	4436.047195		Alco	21.911426
	Age	4289.087958		Active	13.288970
5	Ap_hi	59514.160186	**10**	Ap_hi	59514.160186
	Ap_lo	23057.693730		Ap_lo	23057.693730
	Bmi	4436.047195		Bmi	4436.047195
	Age	4289.087958		Age	4289.087958
	Cholesterol	1226.927745		Cholesterol	1226.927745
6	Ap_hi	59514.160186		Gluc	229.196698
	Ap_lo	23057.693730		Gender	49.914749
	Bmi	4436.047195		Alco	21.911426
	Age	4289.087958		Active	13.288970
	Cholesterol	1226.927745		Smoke	4.430920
	Gluc	229.196698			
7	Ap_hi	59514.160186			
	Ap_lo	23057.693730			
	Bmi	4436.047195			
	Age	4289.087958			
	Cholesterol	1226.927745			
	Gluc	229.196698			
	Gender	49.914749			
8	Ap_hi	59514.160186			
	Ap_lo	23057.693730			
	Bmi	4436.047195			
	Age	4289.087958			
	Cholesterol	1226.927745			

The performances of classifiers were evaluated with all features successively as shown in Table IV. With all selected sets of features, the MLP classifier outperformed all the other models achieving the highest accuracy. Additionally, results showed that the highest accuracy was achieved using the top five features and choosing from six to ten features achieved slightly lower accuracy with only a 0.1% difference. Also, choosing from one to four features achieved lower accuracy with only a 1.0%. The SVM comes after with 86.6% accuracy. DT, RF, and LR almost achieve the same accuracy and NB achieved the lowest accuracy among all models.

Afterwards, the top five features were used as input to the classifiers and performance was evaluated in terms of accuracy. The first test was the LR classifier, which produced an accuracy of 85.5%. We performed the second test of the dataset for the RF classifier and a third test for the DT classifier. These classifiers achieved accuracies of 86% and 85.9%, respectively, and are approximately equal to the first classifier accuracy.

In addition, we used NB and KNN classifiers and achieved accuracies of 83.4% and 84.7%, respectively, which are less than the previous classifiers. Furthermore, we performed tests for the SVM and MLP classifiers; they achieved the highest accuracies of 86.6% and 87.2%, respectively.

We also measured the performance of classifiers for several measurements, to illustrate their robustness and prediction capability. Table VI shows a comparison of classifiers based on several measurements. In terms of the sensitivity measure, the LR classifier had 84.34%. In specificity and precision measures, the SVM classifier reached the highest rates of 95.51% and 96.06%, respectively. Finally, for the F-measure, the MLP classifier had 88.13% and the best accuracy was also achieved by MLP at 87.23%.

To further investigate the models that were selected for cardiovascular disease prediction, we display a ROC curve as shown in Fig. 5. The ROC curve is a metric for each classifier's ability. The model performs best for predicting when the area value is closer to one. It is clear to note that the ROC curve and accuracy result of MLP is the best among the other classifiers for predicting cardiovascular disease.

TABLE VI. EVALUATION PARAMETERS FOR ALL CLASSIFIERS (VALUES LISTED IN PERCENTAGES)

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure
LR	85.54	84.34	87.18	90.01	87.09
RF	86.03	82.19	91.28	92.82	87.19
DT	85.93	81.24	92.34	93.57	86.97
NB	83.38	76.44	92.89	93.65	84.18
KNN	84.56	83.66	85.79	88.97	86.24
SVM	86.63	80.16	95.51	96.06	87.39
MLP	87.23	82.01	94.37	95.23	88.13

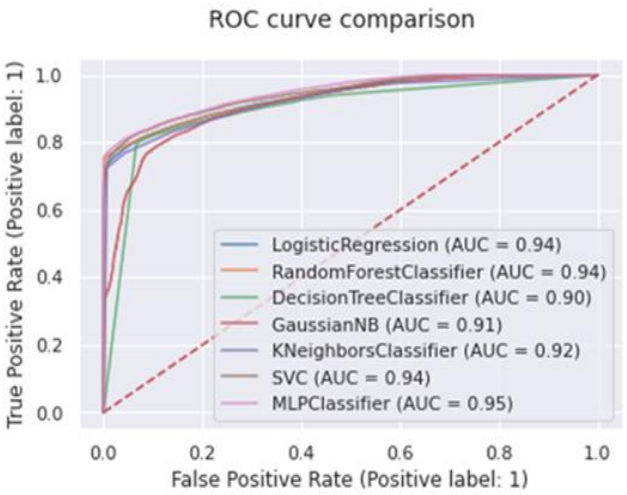


Fig. 5. ROC Curve for all Classifiers.

V. CONCLUSION AND FUTURE WORK

This research studies the performance of machine learning techniques to predict the probability of cardiovascular disease. A dataset of cardiovascular disease for 70000 patients was used for our experiment. Models' performance was evaluated in terms of their accuracies. we have introduced some steps for pre-processing the dataset. In addition, we chose more informative features that impact the performance of the models. Results show that the MLP model demonstrated the highest accuracy in predicting cardiovascular disease.

For our future work, different feature selection techniques can be used to explore the best. More datasets can be used for better and more accurate evaluation. Finally, deep learning techniques can be applied to the prediction problem.

REFERENCES

[1] N. Heart, Lung, B. Institute, N. American, and A. for the S. of Obesity, "The practical guide: identification, evaluation, and treatment of overweight and obesity in adults," Phys. Lett. Sect. A Gen. At. Solid State Phys., vol. 379, no. 10–11, pp. 870–872, 2000, doi: 10.1016/j.physleta.2015.01.006.

[2] D. Deng, P. Jiao, X. Ye, and L. Xia, "An image-based model of the whole human heart with detailed anatomical structure and fiber orientation," Comput. Math. Methods Med., vol. 2012, 2012, doi: 10.1155/2012/891070.

[3] M. Elhneiti and M. Al-Hussami, "Predicting Risk Factors of Heart Disease among Jordanian Patients," Health (Irvine. Calif)., vol. 09, no. 02, pp. 237–251, 2017, doi: 10.4236/health.2017.92016.

[4] "Cardiovascular diseases." https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Nov. 29, 2021).

[5] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 302–305, 2020, doi: 10.1109/ICESC48915.2020.9155586.

[6] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, "Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review," Am. J. Prev. Med., vol. 61, no. 4, pp. 596–605, 2021, doi: 10.1016/j.amepre.2021.04.016.

[7] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," Mater. Today Proc., vol. 37, no. Part 2, pp. 3213–3218, 2020, doi: 10.1016/j.matpr.2020.09.078.

[8] C. J. Harrison and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to natural language processing," BMC Med. Res. Methodol., vol. 21, no. 1, pp. 1–18, 2021, doi: 10.1186/s12874-021-01347-1.

[9] M. Batta, "Machine Learning Algorithms - A Review ," Int. J. Sci. Res. (IJ, vol. 9, no. 1, pp. 381-undefined, 2020, doi: 10.21275/ART20203995.

[10] E. F. Morales and J. H. Zaragoza, "An introduction to reinforcement learning," Decis. Theory Model. Appl. Artif. Intell. Concepts Solut., pp. 63–80, 2011, doi: 10.4018/978-1-60960-165-2.ch004.

[11] M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, "A review of classification problems and algorithms in renewable energy applications," Energies, vol. 9, no. 8, pp. 1–27, 2016, doi: 10.3390/en9080607.

[12] S. Pandey, M. Supriya, and A. Shrivastava, "Data Classification Using Machine Learning Approach," no. June, 2018, doi: 10.1007/978-3-319-68385-0.

[13] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," J. Reliab. Intell. Environ., vol. 7, no. 3, pp. 263–275, 2021, doi: 10.1007/s40860-021-00133-6.

[14] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020, 2020, doi: 10.1109/ic-ETITE47903.2020.242.

[15] J. Vijayashree and H. P. Sultana, "A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier," Program. Comput. Softw., vol. 44, no. 6, pp. 388–397, 2018, doi: 10.1134/S0361768818060129.

[16] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," 2019 2nd Int. Conf. New Trends Comput. Sci. ICTCS 2019 - Proc., pp. 0–5, 2019, doi: 10.1109/ICTCS.2019.8923053.

[17] V. W. Xqdo, "Computational Analysis of Machine Learning Algorithm to predict Heart Disease," vol. 5, pp. 960–964, 2021.

[18] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[19] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mob. Inf. Syst., vol. 2018, 2018, doi: 10.1155/2018/3860146.

[20] G. Renugadevi, G. Asha Priya, B. Dhivyaa Sankari, and R. Gowthamani, "Predicting heart disease using hybrid machine learning model," J. Phys. Conf. Ser., vol. 1916, no. 1, 2021, doi: 10.1088/1742-6596/1916/1/012208.

[21] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart disease prediction using feature selection and ensemble learning techniques," Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021, no. Icicv, pp. 994–998, 2021, doi: 10.1109/ICICV50876.2021.9388482.

[22] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik, and A. U. Zaman, "Heart Disease Prognosis Using Machine Learning Classification Techniques," 2021 6th Int. Conf. Converg. Technol. I2CT 2021, pp. 1–6, 2021, doi: 10.1109/I2CT51068.2021.9418181.

[23] "Cardiovascular Disease dataset | Kaggle." https://www.kaggle.com/sulianova/cardiovascular-disease-dataset (accessed Nov. 29, 2021).

[24] K. J. Bowen, V. K. Sullivan, P. M. Kris-Etherton, and K. S. Petersen, "Nutrition and Cardiovascular Disease—an Update," Curr. Atheroscler. Rep., vol. 20, no. 2, 2018, doi: 10.1007/s11883-018-0704-3.

[25] U.S. Department of Health and Human Services, How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease. 2010.

[26] J. Tan, X. Zhang, W. Wang, P. Yin, X. Guo, and M. Zhou, "Smoking, blood pressure, and cardiovascular disease mortality in a large cohort of chinese men with 15 years follow-up," Int. J. Environ. Res. Public Health, vol. 15, no. 5, 2018, doi: 10.3390/ijerph15051026.

[27] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[28] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, and T. Zhu, "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework," 2017 IEEE 2nd Int. Conf. Big Data Anal. ICBDA 2017, pp. 228–232, 2017, doi: 10.1109/ICBDA.2017.8078813.

[29] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020, vol. 1, no. 6, pp. 177–181, 2020, doi: 10.1109/ICACCCN51052.2020.9362842.

[30] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," Int. J. Eng. Res. Appl., vol. 3, no. 5, pp. 605–610, 2013.

[31] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359–366, 1989, doi: 10.1016/0893-6080(89)90020-8.

A Solution for Automatic Counting and Differentiate Motorcycles and Modified Motorcycles in Remote Area

Indrabayu[1], Intan Sari Areni[2], Anugrayani Bustamin[3]
Elly Warni[4], Sofyan Tandungan[5], Rizka Irianty[6], Najiah Nurul Afifah[7]
Department of Informatics, Universitas Hasanuddin, Makassar, Indonesia[1, 3, 4, 6, 7]
Department of Electrical Engineering, Universitas Hasanuddin, Makassar, Indonesia[2]
College of Informatics Management and Computers STMIK AKBA, Makassar Indonesia[5]

***Abstract*—Motorcycles are the most significant contributor to the vehicle numbers in Indonesia, about 81% of all vehicles in the country. In addition, the growth of modified motorcycles has also increased in several areas, particularly remote places. Many studies have been conducted for detecting vehicles. However, most vehicle detection studies were conducted to detect cars or four-wheeled vehicles, and only a few studies were done to detect motorcycles. Further problems increase if the system is implemented in remote areas with limited electricity power resources that need low-cost budget specification computation. This study detects and calculates the number of motor vehicles and modified motorcycles passed on a highway from video data. It proposed Machine Learning instead of Deep Learning to suit the low computational video in remote areas. Computer vision-based methods used in the prediction are optical flow and Histogram Oriented Gradient (HOG) + Support Vector Machine (SVM). Five videos were used in the system testing, taken from the roadsides using a static camera with a resolution of 160x112 pixels at ±135º angle. This research showed that the accuracy of motorcycles and modified motorcycles detection and calculation systems using the HOG + SVM method is higher than the optical flow method. The average accuracy of HOG + SVM for motorcycles and modified motorcycles is 89.70% and 95.16%, respectively.**

Keywords*—Histogram of oriented gradient; optical flow; vehicles counting; support vector machine

I. INTRODUCTION

The government can observe the density level on the roads by utilizing cameras installed on the corners of the road. However, the utilization of the monitoring camera is still minimal. This technology does not differ from a machine that still requires other parties to operate to make it more useful [1]. Therefore, we need a system that can automatically detect moving vehicles and count vehicles through recorded video from monitoring cameras. However, several areas in southern Sulawesi where many motorcycle and modified motorcycles operate are in the remote area.

Systems that detect and count vehicles in traffic conditions can use active or passive sensors. This research will focus on systems with passive sensor technology because these sensors utilize computer vision for detection that is cheaper than active sensors [2]. Computer vision works by processing image data using combinations of image processing algorithms, artificial intelligence, pattern recognition, and computer graphics to produce information from the image [3].

Several algorithms or methods can be used to estimate the number and types of vehicles, including Optical Flow, Gaussian Mixture Models, Histogram Oriented Gradients (HOG), and Viola-Jones. In this study, the optical flow method is used in the image segmentation process in separating moving objects (vehicles or other moving objects) from stationary objects (roads or other fixed objects) by producing a motion vector that will be thresholded to distinguish objects. This method has been used in various fields, such as facial expression recognition [4], disease detection [5], virtual reality [6], object recognition [7], people counting [8], and gesture recognition [9]. Many previous researchers have used the optical flow in the vehicle recognition field. One of which is Sun et al. [10], who use it to detect and track vehicles in complex traffic conditions with shadow and occlusion. This system combines optical flow and immune particle filter, which increases tracking reliability to work well even in low visibility conditions.

Furthermore, optical flow is also used in [11] in combination with Convolutional Neural Network (CNN). The proposed method was evaluated under challenging environmental conditions. The experimental results showed 96.3% mean detection and 96.8% calculation precision.

In addition to optical flow, the HOG and Support Vector Machine (SVM) methods are also used to compare vehicle detection and counting system accuracy. These methods use the characteristics of the gradient distribution to describe the characteristic shape of an object to recognize the object. The results from HOG features are converted into feature vectors to be processed and trained in the SVM classifier. Finally, in the test stage, objects in the frame will be recognized by comparing the level of similarity of the gradient distribution trained to the gradient distribution in the test image. The combination of HOG and SVM methods has been used for various object detection. In [12], the HOG method extracted wood species and SVM to classify wood species. In [13], the researcher built an automatic mango detector system by combining SVM classifiers trained using HOG features and

image segmentation. Furthermore, HOG and SVM were also used in [14] to detect and classify airborne fungal spores.

Many studies have been conducted in vehicle detection [15],[16]. Some systems are even built to work under challenging conditions, such as dusty weather [17]. However, most vehicle detection studies were conducted on vehicles, and only a few studies were conducted to detect modified motorcycles. In addition, the growth of modified motorcycles or motorized tricycles is now increasingly out of control in some areas in Indonesia [18]. In this study, a comparative analysis of several computer vision algorithms for detecting motorcycles and modified motorcycles will be carried out. This study aims to accurately count the motorcycles and modified motorcycles and find the algorithm's influence on accuracy and processing time.

The structure of this paper is as. In Section 2, the background theory is described. The methodology is explained in Section 3. The result is discussed in Section 4. Finally, Section 5 concludes the paper and discusses future work.

II. Background Theory

A. Traffic Monitoring System

Traffic monitoring systems involve data collection for describing the characteristics of vehicles and their movements on the highway. Most vehicle counting and detection in traffic systems use sensor technology based on radar, microwave, tubes, and loop detectors. Sensors that reflect signals are called active sensors. The active sensor calculates the distance between the source and the target by measuring the time duration between emission and detection of the reflected signal.

On the other hand, sensors that do not reflect signals also can be used in vehicle detection on the highway. This type of sensor is called a passive sensor, an optical sensor that tends to be cheaper because it utilizes cameras and computer vision. This camera-based sensor can extract information more comprehensively, such as vehicle motion, shape, and color. The camera can track passing vehicles and their movement through complex and long-straight roads with precise positioning. A camera-based sensor can be successful if it can be operated in real-time [3].

B. Optical Flow

Optical flow is a visible movement caused by changes in brightness between two images. Optical flow occurs due to the relative movement between the observed object and the observer that can be seen in Fig. 1. The movement captured by optical flow is a movement in a two-dimensional plane [19].

The basic concept of optical flow is to see changes in the brightness of a point in two images. The brightness of a point will be compared with the surrounding area in the following image or frame. From the comparison results, the system can track an object using the brightness of the point that shows the position of the point in the following image or at a different time.

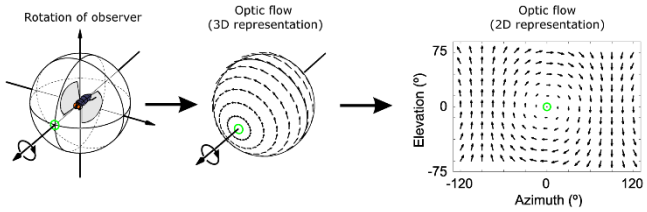


Fig. 1. Optical Flow [19].

Optical flow in computer vision is often used to mark and measure the movement of objects. By observing the intensity or brightness of two sequential images, information on the movement pattern of brightness in the image for each pixel can be obtained. If the pixel intensity value is located on *x, y* at time *t*, its value would be the same as the pixel located *(x+δx,y+ δy)* at time *t+ δt*.

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \quad (1)$$

By applying the *Taylor series* to the right-hand side of eq. (1), eq. (2) can be obtained:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\delta I}{\delta x}\delta x + \frac{\delta I}{\delta y}\delta y + \frac{\delta I}{\delta t}\delta t + H.O.T \quad (2)$$

By ignoring the higher-order terms (H.O.T.) after simplifying the notation, a simple form of optical flow is obtained, where *u* and *v* are flow vectors for each pixel, I_x and I_y are spatial gradients of brightness intensity, and I_t is the derivative of brightness intensity concerning time.

$$I_x u + I_y v + I_t = 0 \quad (3)$$

C. Blob Analysis

Blob analysis is a technique used to express the pixel area of an image that becomes the focus of detection. This technique is used in optical flow to detect vehicles. The determination of the blob area for each object in the foreground segmentation process needs to be analyzed because the blob value for each object is different that influenced by object features such as size, type, and techniques in obtaining video data.

D. Histogram of Oriented Gradient (HOG)

Histogram of Oriented Gradient (HOG) is used in image processing for object detection purposes. This technique calculates the gradient value in a specific area of an image. Each image has characteristics indicated by a gradient distribution. These characteristics are obtained by dividing the image into small areas called cells. Each cell is composed of the histogram of gradients. Combining these histograms is used as a descriptor that represents an object.

HOG works using the window shift concept by calculating the gradient vector obtained for each window. The value and direction of the gradient vector in a particular area will display the characteristics of the gradient distribution of an image. Gradient distribution characteristics will describe the shape of an object in the image so that training can be carried out to recognize an object. Finally, objects in an image will be recognized by comparing the level of similarity of the gradient

distribution in the trained images to the gradient distribution in the target image. The results of the HOG feature are converted into feature vectors to be processed in the Support Vector Machine classifier.

E. Support Vector Machine

Supervised learning is a learning process where the training data has been labeled according to their respective class before the training begins. Furthermore, the system only checks the similarity of the features between the new incoming data and the training data, then labels them according to the most similar training data. One of the supervised learning methods is Support Vector Machines (SVM). The SVM algorithm aims to find a hyperplane that can separate classes with the maximum distance (margin/gap) between a particular class's borderlines (support vectors) and the borderlines of other classes. The basic idea of SVM itself is to find a linear decision surface (hyperplane) or barrier that separates one class from another with the most significant distance/gap/margin [6].

F. Tracking

Searching for moving objects in a sequence of frames is known as tracking. Tracking can be done by using object feature extraction and detecting moving objects/objects in the frame sequence. In computer vision, object tracking is a process that aims to follow the movement of an object. Furthermore, tracking also can be used for counting the detected vehicles.

III. Methodology

The block diagram of this research is depicted in Fig. 2. Following the input image, there is a preprocessing consisting of two stages. Subsequently, two detection algorithms will be reviewed in each model, namely Optical Flow and Histogram Oriented Gradients-Support Vector Machine (HOG-SVM). The results from the previous process will be used to calculate the number of vehicles that have been identified.

A. Input

Data acquisition was carried out on Jalan Adhyaksa Baru, Makassar, by placing a camera 4 meters above the ground using an iron pole, as shown in Fig. 3. The camera was placed on an iron pole at an angle of ±135º. This angle was used so that the passing vehicle could be seen or caught clearly by the camera. Fig. 3 illustrates the camera's position on the iron pole when collecting the data. The pole is attached to a power pole on the side of the road. Data is collected in the form of video. In this research, the data was taken from the rear corner of the vehicle where the recorded vehicles moved away from the camera.

This study designed two detection models, i.e., motorcycles and modified motorcycles. A Modified motorcycle is two-wheeled vehicles that have changed from their basic form intended to increase passenger capacity. A physical comparison of these two vehicles can be seen in Fig. 4.

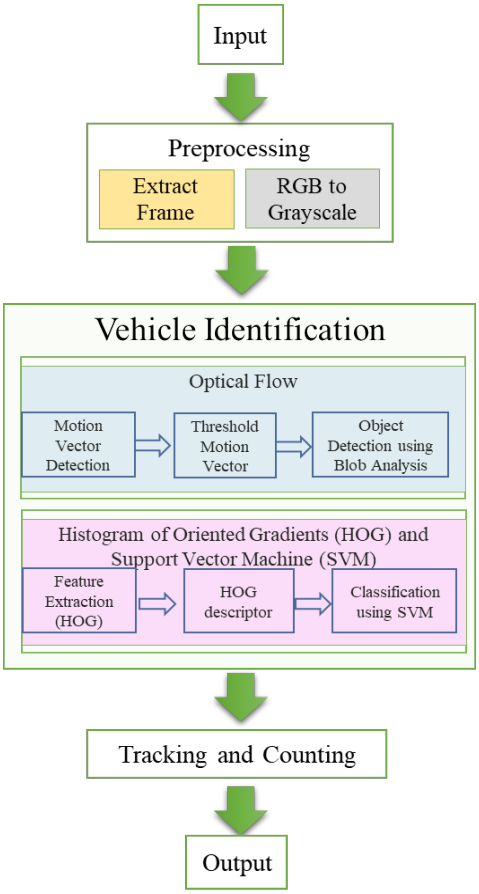


Fig. 2. Block Diagram of Vehicle Identification.

Fig. 3. Illustration of Data Acquisition Process.

(a) (b)

Fig. 4. Sample of Dataset (a) Motorcycle and (b) Modified Motorcycle.

B. Preprocessing

After the input data from the video is obtained, the preprocessing stage will be carried out. Preprocessing is a stage for preparing the data before further processing to get maximum results. This stage will convert RGB frames to grayscale as an input for the next stage.

C. Vehicle Identification

This stage aims to identify the vehicle by implementing two different algorithms, which are optical flow and HOG + SVM classification.

1) Optical Flow method. This algorithm comprises three essential steps as follows:

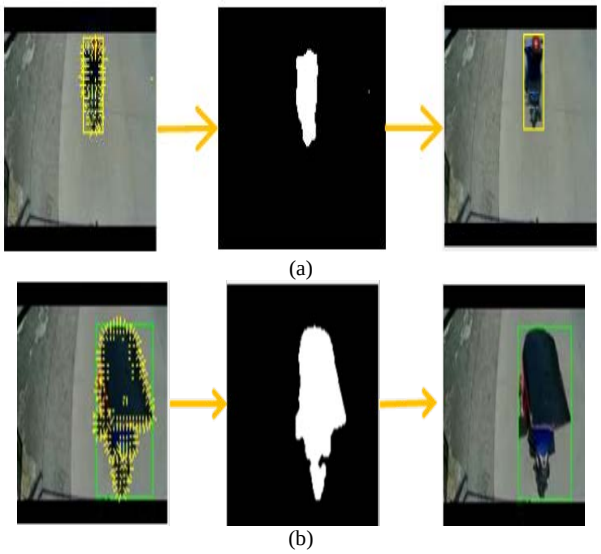
a) Motion vector detection: This stage will detect vehicle movement by calculating the estimated optical flow. This process uses input in the form of grayscale frames. A yellow box around the moving object will indicate the detected motion vector.

b) Motion vector thresholding: Detected motion vectors will be subjected to a thresholding process to produce binary format frames, where foreground objects are labeled as one, and background objects are labeled as zero.

c) Objects detection using blob analysis: The thresholding process produces blobs of objects that the blob algorithm will analyze. The blob area size used in this study is 500-1000 pixels for motorcycles and 2100-2700 pixels for modified motorcycles.

Fig. 5 shows the steps of the identification process for motorcycles and modified motorcycle objects, which consist of the motion vector detection process, thresholding, and object detection with blob analysis.

2) HOG+SVM Method. This method starts by extracting features from the image that has been obtained in the previous process in preparation for the training phase. The HOG+SVM method comprises the following stages.



(a)

(b)

Fig. 5. Detection Process in Optical Flow (a) Motorcycle and (b) Modified Motorcycle.

a) Training process: This stage aims to train the system for object classification (motorcycles and modified motorcycles) to generate a model used in the testing stage. This stage is carried out with several positive and negative training and negative data.

For the Motorcycle Detection Model, positive data consists of positive data, namely motorcycle images, and negative data, such as images of cars, roads, and pedestrians. At the training stage, the Modified Motorcycle Detection Model uses positive data in the form of modified motorcycle images and negative data in images of cars, roads, and other object images.

For the Motorcycle Detection Model, the training data includes 51 positive data and 300 negative data of motorcycles, and Modified Motorcycle Detection Model uses training data that consist of 100 positive data and 363 negative data of modified motorcycles.

After preparing each model's training data, the feature extraction process executes the HOG function. The output is the vector of the HOG descriptor, as shown in Fig. 6.

After the HOG extraction, classification was carried out using the SVM algorithm. This method classifies or separates motorcycles and modified motorcycles objects from other objects in the scenes. The vector obtained in the training process will be stored as a model and used for detection in the testing process.

b) Detection: A shift or sliding window is used to detect the feature descriptor in this stage. The sliding window will iterate 5 pixels per frame with a window size of 62x62 pixels for motorcycles detection and 71x71 pixels for modified motorcycles detection. HOG extraction is performed in each sliding window. The results of the HOG extraction will be classified using the SVM classifier generated at the training stage. The illustration of the detection process can be seen in Fig. 7.

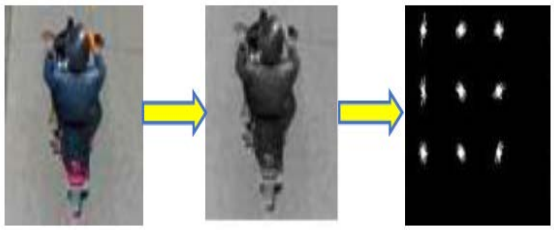


Fig. 6. Illustration of the HOG Algorithm.

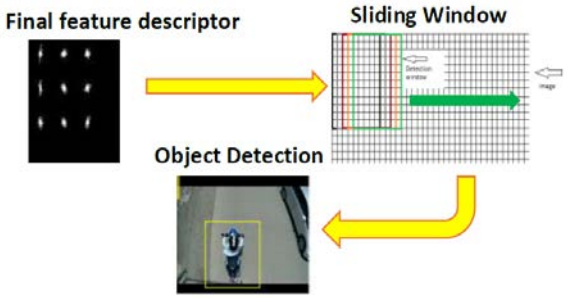


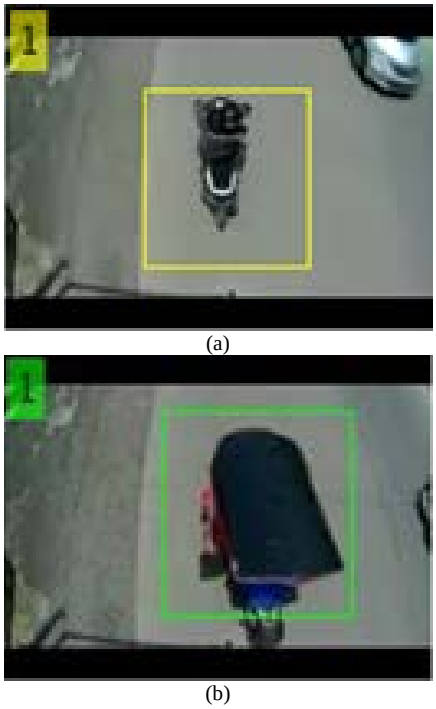
Fig. 7. HOG+SVM Method Object Detection Process.

D. Vehicle Tracking and Counting

The vehicle count will be based on the number of tracks generated by the tracking algorithm. This system uses the Kanade Lucas Tomasi Tracking (KLT) algorithm to count. The KLT tracking involves information from the previous frame, where the detected vehicle and the previous frame are input to this process. The output of this process is the vehicle's position in the current frame. The information of the current frame will be used as a tracking reference for the next frame. The tracking process will continue until the end of the frame. The KLT tracking is based on the feature values comparison in the frames based on a score of each bounding box.

E. Output

The output of this system is videos composed of a collection of detected frames with RGB data type. Each frame will show the number of detected vehicles in the upper left corner of the frame as shown in Fig. 8.



(a) (b)

Fig. 8. Output Display System: (a). Motorcycle Detection , (b) Modified Motorcycle.

F. System Performance Analysis

The confusion matrix is used to measure the detection system performance, and the results are shown in Table I.

The system accuracy for each video is calculated using Eq. (4).

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (4)$$

TABLE I. CONFUSION MATRIX OF THE DETECTION RESULTS

	Predicted	***Not predicted***
Positive Data	*True Positive* (*tp*)	*False Positive* (*fp*)
Negative Data	*False Negative* (*fn*)	*True Negative* (*tn*)

where:

- True Positive (tp), the motorcycles or modified motorcycles objects detected by the bbox
- True Negative (tn), not motorcycles object or modified motorcycles that is not detected by bbox.
- False Positive (fp), not motorcycles objects or modified motorcycles but detected by bbox.
- False Negative (fn), motorcycles object or modified motorcycles that is not detected by bbox.

IV. RESULTS AND DISCUSSION

The average results from both the optical flow and HOG+SVM methods are presented in Fig. 9. Five videos are used as the test data for detecting and counting the modified motorcycles and motorcycles, each of which is 1 minute long. The testing stage is carried out using the optical flow and HOG+SVM methods. The first test is accuracy analysis using the optical flow method. Table II shows the accuracy obtained using optical flow. The next test is the accuracy analysis using the HOG + SVM method. Based on (4), the measurement of system performance in this study focused on the TP value and considered the TN, FP, and FN values. The TP value for each video is relatively similar, but the FP value obtained tends to fluctuate, causing a decrease in accuracy. Table III shows the results of accurate measurement for the HOG + SVM method.

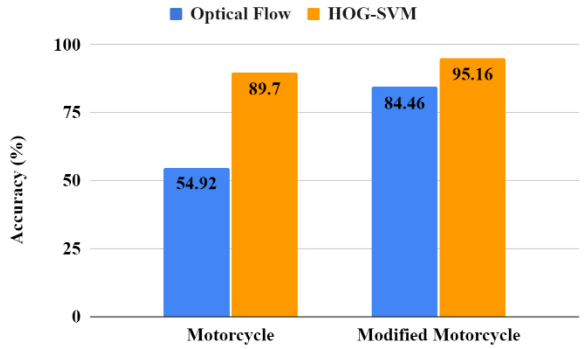


Fig. 9. Average Accuracy of the System.

TABLE II. SYSTEM ACCURACY USING OPTICAL FLOW

Model	**Video**	**TP**	**TN**	**FP**	**FN**	**Accuracy (%)**	**Average Accuracy (%)**
Motorcycle	1	20	3	13	0	63.89	54.92
	2	19	3	11	1	64.71	
	3	13	2	16	0	48.39	
	4	18	3	20	2	48.84	
	5	17	3	21	0	48.78	
Modified Motorcycle	1	2	33	5	0	87.50	84.46
	2	1	43	3	0	93.62	
	3	2	27	2	0	93.55	
	4	1	18	11	1	61.29	
	5	1	37	6	0	86.36	

TABLE III. SYSTEM ACCURACY USING HOG-SVM

Model	Video	TP	TN	FP	FN	Accuracy (%)	Average Accuracy (%)
Motorcycle	1	21	11	1	3	88.89	89.70
	2	20	10	1	3	88.24	
	3	11	17	2	1	90.32	
	4	21	17	2	3	88.37	
	5	18	20	1	2	92.68	
Modified Motorcycle	1	1	38	0	1	97.50	95.16
	2	1	44	2	0	95.74	
	3	1	28	1	1	93.55	
	4	1	28	1	1	93.55	
	5	0	42	1	1	95.45	

Optical flow accuracy is 54.92% and 84.46% consecutively for motorcycles and modified motorcycles. In comparison, HOG+SVM performs 89.70% for motorcycles and 95.16% for modified motorcycles. The results show that HOG+SVM outperforms Optical flow in terms of accuracy.

To measure the performance of the system, data from the detection results for both motorcycle (in Fig. 10) and modified motor (in Fig. 11) are categorized according to the confusion matrix variable.

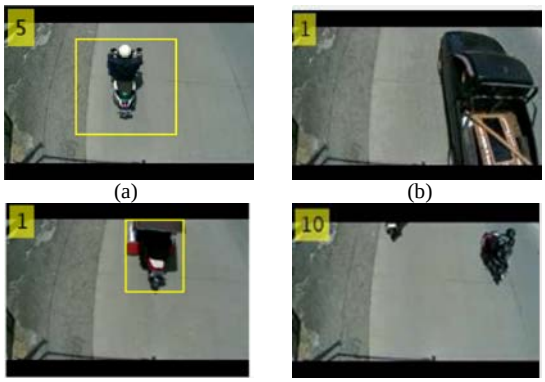


Fig. 10. Motorcycle Detection Results has Categorized in Confusion Matrix: (a) True Positive, (b) True Negative, (c) False Positive and (d) False Negative.

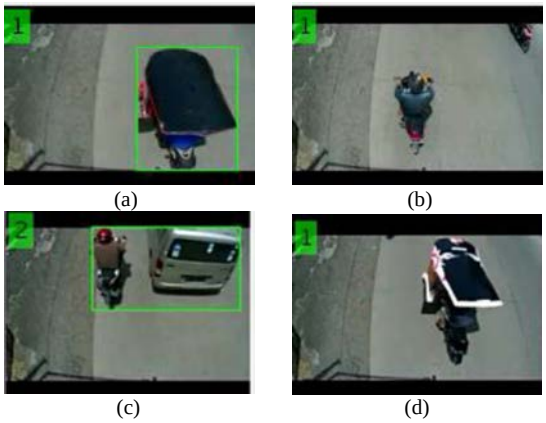


Fig. 11. Modified Motorcycle Detection Results has Categorized in Confusion Matrix: (a) True Positive, (b) True Negative, (c) False Positive and (d) False Negative.

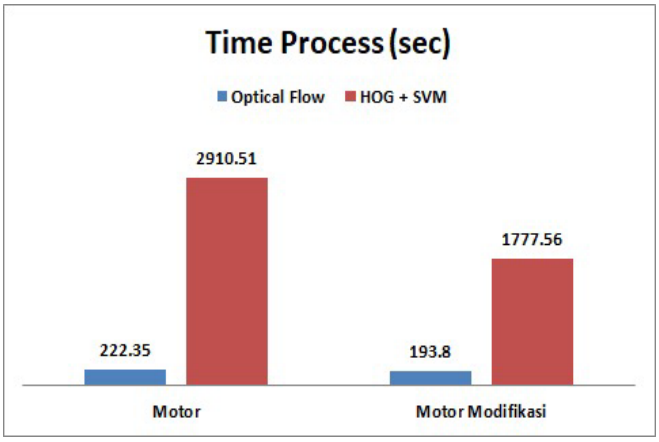


Fig. 12. Time Process of the Systems.

Fig. 12 shows the average system time, representing the processing time to detect and count the motorcycles and modified motorcycles in the video for each method. The accuracy measurement for the optical flow method is shown in the results of 54.92% accuracy for motorcycles and 84.46% for modified motorcycles. On the other hand, the HOG+SVM method produced average accuracy of 89.70% for motorcycles and 95.16% for modified motorcycles. Based on the results, the HOG+SVM method obtained higher accuracy than the optical flow method.

However, the optical flow method took less time to detect, which means the optical flow method is faster than HOG+SVM due to optical flow detected vehicles based on the blob area that had been determined in a frame, while the HOG+SVM method detected vehicles by checking the HOG feature through a window shift. HOG+SVM checked one window at a time on a frame. Thus, the HOG+SVM time process is slower than the optical flow.

V. CONCLUSION

This research aimed to design motorcycles and modified motorcycles classification and automatic counting systems through video data. The system test results showed that motorcycles and modified motorcycle detection accuracy using the HOG+SVM method was higher than the optical flow method. The average accuracy of HOG+SVM for motorcycles was 89.70%, and modified motorcycles was 95.16%, compared to the optical flow method with 54.92% accuracy for 84.46% for modified motorcycles. However, the computational time required for the optical flow method was faster than the HOG+SVM method, which were 222.35s for motorcycles and 193.80s for modified motorcycles. Meanwhile, HOG+SVM took 2910.51s and 1777.56s for motorcycles and modified motorcycles (*bentor* in Indonesia Language), respectively.

This research can be developed with variations of other vehicles in terms of categories and image specifications. Data retrieval in this study is taken in the daytime at a specific time only; hence further development for night conditions is needed before it can be used to support Intelligent Transport System technology in the real-world environment.

ACKNOWLEDGMENT

This work is supported by the LPPM - Universitas Hasanuddin (UNHAS) via PDUPT-Ristekdikti Grant 2021 with contract number 752/UN.4.22/PT.02.00/2021, Indonesia and AIMP Research Group.

REFERENCES

[1] Jusman, "CCTV and Smart Governments," *The Fajar News*, March 14, 2016. [Online]. Available: http://makassarkota.go.id/berita-1105-cctv-dan-pemerintahan-yang-cerdas.html. [Accessed December 20, 2021] (In Indonesian).

[2] S. Jung, Y. Cho, D. Kim, and M. Chang, "Moving Object Detection from Moving Camera Image Sequences Using an Inertial Measurement Unit Sensor," *Applied Sciences*, vol. 10, no. 1, 268, 2020.

[3] V. Wiley and T. Lucas, "Computer Vision and Image Processing: A Paper Review," *International Journal of Artificial Intelligence Research*, vol. 2, no. 1, pp. 28-36, 2018.

[4] G. Patil and P. Suja, "Emotion recognition from 3D videos using optical flow method," in *2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon)*, India, 2017, pp. 825-829.

[5] M. Abdel-Nasser, A. Moreno, H. A. Rashwan, D. Puig, "Analyzing the evolution of breast tumors through flow fields and strain tensors," *Pattern Recognit. Lett.*, vol. 93, pp. 162–171, 2017.

[6] G. Ren, W. Li, E. O'Neill, "Towards the design of effective freehand gestural interaction for interactive TV," *J. Intell. Fuzzy Syst.*, vol. 31, no. 5, pp. 2659–2674, 2016.

[7] S. Gujunoori and M. Oruganti, "Tracking and Size Estimation of Objects in Motion using Optical flow and K-means Clustering," in *2nd International Conference On Emerging Computation and Information Technologies (ICECIT)*, India, 2017, pp. 1-6.

[8] Tokta, Aybars and A. K. Hocaoglu, "A Fast People Counting Method Based on Optical Flow," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Turkey, 2018, pp. 1-4.

[9] M. Wrzalik and D. Krechel, "Human Action Recognition Using Optical Flow and Convolutional Neural Networks," in *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Mexico, 2017, pp. 801-805, doi: 10.1109/ICMLA.2017.00-59.

[10] Wei Sun, Min Sun, Xiaorui Zhang, and Mian Li, "Moving Vehicle Detection and Tracking Based on Optical Flow Method and Immune Particle Filter Under Complex Transportation Environments," *Complexity*, vol. 2020, 2020.

[11] A. Gomaa, M. M. Abdelwahab, M. Abo-Zahhad, T. Minematsu, and R. Taniguchi, "Robust Vehicle Detection and Counting Algorithm Employing a Convolution Neural Network and Optical Flow," *Sensors*, vol. 19, no. 20, 4588, 2019.

[12] B. Sugiarto et al., "Wood Identification Based on Histogram of Oriented Gradient (HOG) Feature and Support Vector Machine (SVM) Classifier," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Indonesia, 2017, pp. 337-341, doi: 10.1109/ICITISEE.2017.8285523.

[13] M. J. C. Baculo and N. Marcos, "Automatic Mango Detection using Image Processing and HOG-SVM," in *Proceedings of the 2018 VII International Conference on Network, Communication and Computing (ICNCC 2018)*, Association for Computing Machinery, New York, 2018, pp. 211–215.

[14] M. W. Tahir, N. A. Zaidi, R. Blank, P. P. Vinayaka, M. J. Vellekoop, and W. Lang, "Detection of Fungus Through an Optical Sensor System Using The Histogram of Oriented gradients", in *IEEE Sensors 2016*, Florida, pp. 1-3, 2016.

[15] K.V. Sakhare, T. Tewari, V. Vyas, "Review of Vehicle Detection Systems in Advanced Driver Assistant Systems," *Arch. Computat. Methods Eng.*, vol. 27, pp. 591–610, 2020.

[16] A. F. Abbas, U. U. sheikh, F. T. Al-dhief, and M. N. H. Mohd, "A comprehensive review of vehicle detection using computer vision," *TELKOMNIKA*, vol. 19, no. 3, pp. 838-850, 2021.

[17] Nastaran Yaghoobi Ershadi and José Manuel Menéndez, "Vehicle Tracking and Counting System in Dusty Weather with Vibrating Camera Conditions," *Journal of Sensors*, vol. 2017, Article ID 3812301, 2017.

[18] F. M. Azzam, "Motorcycle Modification is Growing Uncontrollably, Makassar People Representative Council Asking for Government Intervention" *The PosKota*, May 3, 2021. [Online]. Available: https://sulsel.poskota.co.id/2021/05/03/bentor-menjamur-dprd-makassar-minta-pemkot-makassar-lakukan-evaluasi [Accessed December 20, 2021] (In Indonesian).

[19] S. Ranganatha and Y. P. Gowramma, "A Comprehensive Survey of Algorithms for Face Tracking in different Background Video Sequence," *International Journal of Computer Applications*, vol. 181, no. 27, pp. 43-49, 2018.

PAD: A Pancreatic Cancer Detection based on Extracted Medical Data through Ensemble Methods in Machine Learning

Santosh Reddy P*[1]
Research Scholar
Department of CSE
Presidency University, Bengaluru

Chandrasekar M[2]
Associate Professor
Department of CSE
Presidency University, Bengaluru

***Abstract*—The considerable research into medical health systems is allowing computing systems to develop with the most cutting-edge innovations. These developments are paving the way for more efficient medical system implementations, including automatic identification of health-related disorders. The most important health research is being done to predict cancer, which can take several forms and affect many parts of the body. One of the most prevalent tumors that is expected to be incurable is pancreatic cancer. Pancreatic cancer is one of the most common cancers that is projected to be incurable. Previous research has found that a panel of three protein biomarkers (LYVE1, REG1A, and TFF1) found in urine can help detect respectable PDAC. To improve this panel in this study by replacing REG1A with REG1B from extracted data sets into CSV format. Finally, will analyze four significant biomarkers that are found in urine, creatinine, LYVE1, REG1B, and TFF1. Creatinine is a protein that is commonly utilized as a kidney function indicator. Lymphatic vessel endothelial hyaluronan receptor 1 (YVLE1) is a protein that may help tumors spread. REG1B is a protein that has been linked to pancreatic regeneration, while TFF1 is trefoil factor 1, which has been linked to urinary tract regeneration and repair It's impossible to treat it properly once it's been diagnosed. Machine learning and neural networks are now showing promise for accurate pancreatic picture segmentation in real time for early diagnosis. This research looks at how to analyze pancreatic tumors using ensemble approaches in machine learning. According to preliminary data, the proposed technique looks to improve the classifier's performance for early diagnosis of pancreatic cancer.**

***Keywords*—Pancreatic; PDAC; LYVE1; REG1A; TFF1; CA19_9**

I. Introduction

Cancer, according to medical health news analysis, is one of the most troublesome diseases that can appear to be invincible at times. It's possible that it's a hereditary disease because it's caused by abnormalities in genes that control how cells in the human body work. These genetic alterations might be passed down through generations or caused by a person's lifestyle. It is an important organ of the human body, has internal and external secretory functions and is disposed to various diseases. Surgical right of entry and for which prebiopsy is repeatedly impossible [1-5]. Pancreatic cancer is the fourth majority common source of cancer death and the second most important cause of death from neoplasm's disturbing the digestive coordination.

However, regular segmentation of the pancreas remainders a dispute for the subsequent reasons: 1) low soft tissue contrast on CT images. 2) Huge anatomical variations. The pancreas shows great anatomical unpredictability in terms of size and location in the abdominal cavity of patients [6][7]. the pancreas is a deformable yielding tissue. Consequently, the outline and manifestation of the pancreas have great differences in dissimilar individuals. PDAC (pancreatic ductal adenocarcinoma) is a particularly lethal form of pancreatic cancer. The five-year survival rate is less than 10% once diagnosed. However, if the cancer is caught early enough, when tumours are still small and manageable, 5-year survival rates can reach 70%. Unfortunately, many cases of pancreatic cancer go undetected until the disease has progressed throughout the body. As a result, a diagnostic test to detect pancreatic cancer patients could be quite beneficial. Traditionally, blood has been the primary source of biomarkers, however urine is a viable alternative biological fluid. It enables non-invasive sample, high-volume collection, and repeated measurements with ease. There are currently no reliable biomarkers for detecting PDAC earlier. Serum CA19-9, the only biomarker utilized in clinical practice, is not specific or sensitive enough for screening and is primarily employed as a prognostic marker and for monitoring treatment response.

Even though to collects invasive samples, he increases cancer diagnosis when combined with other urine indicators in a study. Previous research has found that a panel of three protein biomarkers (LYVE1, REG1A, and TFF1) found in urine can help detect significant PDAC. We improved this panel in this study by replacing REG1A with REG1B. Finally, we will analyze four significant biomarkers that are found in urine: creatinine, LYVE1, REG1B, and TFF1. Creatinine is a protein that is commonly utilized as a kidney function indicator. Lymphatic vessel endothelial hyaluronan receptor 1 (YVLE1) is a protein that may help tumors spread. REG1B is a protein that has been linked to pancreatic regeneration, while TFF1 is trefoil factor 1, which has been linked to urinary tract regeneration and repair.

*Corresponding Author

II. Literature Survey

The concept of regular automation algorithms and suggest that Support Vector Machine (SVM) is an authoritative classification process for classifying data related to the calculation of Wisconsin Breast Cancer data with a minor proportion of time [8]. Proportion of relative results in stipulations of effectiveness and effectiveness of four algorithms of differences in data retrieval and automatic automation. Initiates a new functioning favor of the medical health system with the intention of predicts the outcome of an average patient in the examination of electronic medical proceedings and the recognized parameters of parameters established for proper functioning [9]. The efficient prognostic data is normally provided by the application coordination with the estimate of data for variable effects, types of effects and the threshold parameter to identify the diagnosis of the disease. Corresponding the medical proceedings, they use and cover for blood cancer, heart failure, diabetes [10].

To develop a function based on the red convolution neuronal representation to analyze rectal prescribed amount sharing and predict rectal toxicity in patients with uterine cancer, by means of data as of combined radiotherapy (EBRT) and brachytherapy (BT) [11]. They adopted is a somewhere to live and transfer strategy to influence patient data. The adaptive synthetic model technique is used to increase the dates for footage data losses and loss factors. Produce Gradient Activation Weight Map (Grad-CAM) classes to generate RSDM discriminate regions with the calculate model. The CNN-based representation for predicting rectal dose by means of transfer therapy for uterine cancer radiotherapy is analyzed by means of a conjunction of experimental outcome [12].

III. Methods and Materials

Neural Designer was used to tackle this problem. You can utilise the trial to follow it step by step. Because the variable to be predicted is categorical, this is a classification project (no pancreatic disease, benign hepatobiliary disease, or pancreatic cancer). The goal is predicting the presence of disease before it's diagnosed, and more specifically, differentiating between pancreatic cancer versus non-cancerous pancreas condition and healthy condition

A. Data Set

Barts Pancreas Tissue Bank, University College London, University of Liverpool, Spanish National Cancer Research Centre, Cambridge University Hospital, and University of Belgrade all contributed to the data collection. A total of 590 urine samples were tested for the biomarker panel, including 183 control samples, 208 benign hepatobiliary disease samples (of which 119 were chronic pancreatitis), and 199 PDAC samples. Data source, Variables, Instances, and Missing values are the four concepts that make up this system. The information used to generate the model is contained in the data file pancreatic-cancer.csv. There are 509 rows and 14 columns in all. The rows represent the study samples, while the columns representvarious cancer risk variables.

This data collection makes use of the following 16 variables: id of the sample. Each subject is identified by a unique string called a cohort. Cohort 1 samples has been used previously. Cohort 2 samples have been added, with the following sample origin: BPTB: Barts Pancreas Tissue Bank, London, UK; ESP: Spanish National Cancer Research Centre, Madrid, Spain; LIV: Liverpool University, UK; UCL: University College London, UK; BPTB: Barts Pancreas Tissue Bank, London, UK; BPTB: Barts Pancreas Tissue Bank, London, UK; CA 19–9 monoclonal antibody levels in blood plasma, which are frequently elevated in pancreatic cancer patients. Only 350 participants were analysed (one goal of the study was to compare different CA 19-9 cut points from a blood sample to a model built using urine samples), creatinine: A urinary biomarker of renal function. LYVE1: Lymphatic vessel endothelial hyaluronan receptor 1 is a protein discovered in the urine that may have a role in tumour spread. REG1A: Urinary levels of a protein connected to pancreatic regeneration, REG1B: Urinary levels of a protein linked to pancreatic regeneration, REG1B: Urinary levels of a protein linked to pancreatic regeneration TFF1: Only 306 patients had their urinary Trefoil Factor 1 levels evaluated, which could be linked to urinary tract regeneration and repair (one purpose of the study was to assess REG1B vs. REG1A). 3 = Pancreatic ductal adenocarcinoma; 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis) i.e., pancreatic cancer, benign sample diagnosis: Stage: For those who have been diagnosed with a benign, non-cancerous condition, stage: IA, IB, IIA, IIIB, III, IV are the stages of pancreatic cancer. There are a few input variables that must be marked as unused among all of them. Specifically, ‘sample id', which Neural Designer does automatically, ‘sample origin', which only specifies the origin of the patient samples and should not affect the final diagnosis, ‘stage,' which is a variable that only exists for people we already know have cancer, 'patient cohort,' which does not contribute to the final sample diagnosis, and 'benign sample diagnosis,' which is a variable that does not contribute to the final sample diagnosis. The variable corresponding to the biomarker REG1A is not in all the samples of the study. For that reason, we choose to set it as unused too. This decision will not mean a deterioration of the model as the biomarker REG1B improve the results. Once the data set is configured, we can calculate the data distribution of the variables. The following figure depicts the number of patients who have cancer and those who do not. The minimum frequency is 31.0169%, which corresponds to no pancreatic disease diagnosis. The maximum frequency is 35.2542%, which corresponds to benign hepatobiliary disease diagnosis. As we can see, all the samples are well distributed between the three cases.

There should be a partition our dataset into four subsets to compare the accuracy and AUC (Area Under Curve) calculated in this study with those in the paper listed in the references section. Control samples vs. PDAC stages I and II: We only chose healthy person samples and pancreatic cancer stages I and II samples from the raw dataset. Control samples vs. PDAC stages III and IV: Only healthy individual samples and pancreatic cancer stages III and IV samples were chosen from the raw dataset. Benign hepatobiliary disorders vs. PDAC stages I and II: We selected individuals with benign tumour samples and pancreatitis cancer stages I and II samples from the raw dataset. PDAC stages III and IV vs. benign hepatobiliary diseases: Only individuals with benign tumour samples and

pancreatic cancer stages I and II are chosen from the raw dataset. In all these scenarios, the examples are separated into training and testing subsets, with each subset having half of the samples.

B. Stage I & Stage II with Sample Data

The inputs-target correlations of all the inputs with the target are shown in the Fig. 1. This allows us to see how different inputs affect the default.

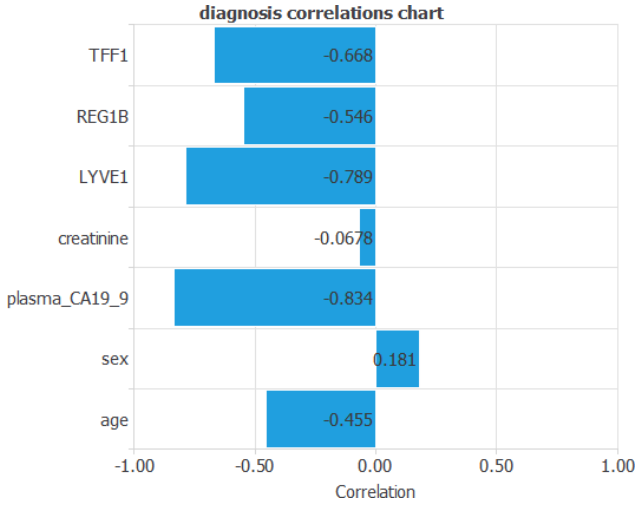


Fig. 1. Diagnosis Correlation Chart for Stage 1 & 2.

The biomarkers LYVE1 and plasma CA19 9 are the most highly associated variables.

C. Stage III & Stage IV with Sample Data

Fig. 2 shows inputs-target correlations of all the inputs with the biomarkers LYVE1 and TFF1 are the most highly associated variables.

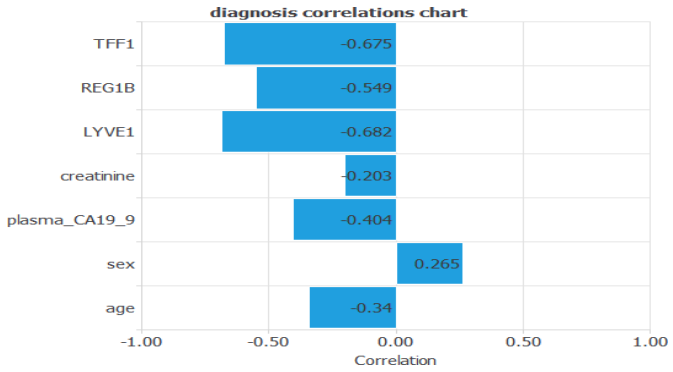


Fig. 2. Diagnosis Correlation Chart for Stage 3 & 4.

D. Implementation

The above Fig. 3 shows the system Design. IPancreatic cancer is one of the most devastating types of cancer, with something like a terrible prognosis in the present environment. Because of its complex visual appearance and indistinct curvature, the pancreas border line is difficult to distinguish from its anatomies in CT/MRI scans. Most relevant health research is available on cancer prediction, which comes in a variety of forms and can affect different sections of the body. Pancreatic cancer is one of the most common cancers that is projected to be incurable. Once diagnosed, it cannot be treated adequately. Machine learning and neural networks are providing promising findings for accurate pancreatic picture segmentation in real time early detection these days. Pancreatic cancer can be classified into five stages. The size and location of the tumour, as well as whether the cancer has spread to the liver, lungs, or abdominal cavity, will determine your diagnosis. It's possible that it's spread to nearby organs, tissues, or lymph nodes. Make sure to discuss your case with your healthcare practitioner. Understanding your pancreatic cancer prognosis might assist you in making an informed treatment selection. According to previous studies, a panel of three protein biomarkers present in urine (LYVE1, REG1A, and TFF1) can assist detect significant PDAC.

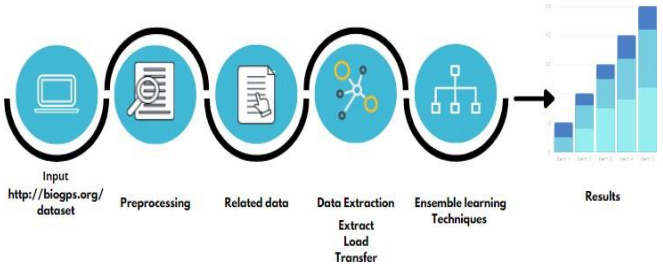


Fig. 3. System Design.

We improved this panel in this study by replacing REG1A with REG1B. Finally, we will analyse four significant biomarkers that are found in urine: creatinine, LYVE1, REG1B, and TFF1. Creatinine is a protein that is commonly utilised as a kidney function indicator. Lymphatic vessel endothelial hyaluronan receptor 1 (YVLE1) is a protein that may help tumours spread. REG1B is a protein that has been linked to pancreatic regeneration, while TFF1 is trefoil factor 1, which has been linked to urinary tract regeneration and repair It's impossible to treat it properly once it's been diagnosed. Machine learning and neural networks are now showing promise for accurate pancreatic picture segmentation in real time for early diagnosis.

1) Naive bayes: To make it easier to understand, I'll go over the theory behind Naive Bayes first, and then use an example to clarify the notions. The Bayes Theorem, which asserts the following equation, inspired the Naive Bayes Classifier.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Rewrite the equation using X (input variables) and y (output variables) to make it easier to understand (output variable). In plain English, this equation calculates the probability of y given input attributes X.

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

We may rewrite P(X|y) as follows because of the naive assumption (therefore the name) that variables are independent given the class.

$P(X|y) = P(X_1|y)* P(X_2|y)*\ldots\ldots\ldots* P(X_n|y)$

Also, because we're solving for y, P(X) is a constant, so we can drop it from the equation and replace it with a proportionality. As a result, we arrive to the following equation.

$$P(y|X) \propto P(X|y) * P(y)$$

Or

$$P(y|X) \propto P(y) * \pi_{i=1}^{n} P(x_i|y)$$

The purpose of Naive Bayes is to choose the classy with the highest probability now that we've reached at this equation. Argmax is a simple operation that finds the argument that gives the target function's maximum value. In this situation, we're looking for the highest y value.

2) Bagging & boosting: When estimating a numerical outcome, aggregating, and voting with a plurality when predicting a class, BAGGING (Fig. 4) is the process of applying Bootstrap sampling on the training dataset, aggregating when estimating a numerical outcome, and voting with a plurality when predicting a class. Bagging, on the other hand, would degrade the performance of stable algorithms such as k-nearest neighbours discriminant analysis, and Nave bayes, because this algorithm uses initial samples that contain about 63 percent of the original data, meaning that each sample is missing about 37 percent of the original data. The Boosting strategy works by combining numerous simple learning algorithms instead of employing a very accurate prediction rule. The update approach then combines all these weak rules to reduce variations and deviations in the individual model rules, leading to a single prediction rule that is significantly more accurate than any of the weak rules alone. There are two main techniques for effectively applying the reinforcement algorithm.

Test error is the minuscule proportion of errors on a recently sampled test set. CT scans can be used to detect if cancer is present and has spread, as well as to guide a biopsy, and can be used to diagnose pancreatic cancer utilizing a variety of imaging modalities. MRIs are used when CT scans aren't a possibility or other tests aren't conclusive. An endoscope can be used to perform ultrasounds from outside the abdomen or through the digestive tract. Why is it so common for pancreatic cancer to be found so late? Because the pancreas is placed deep within the abdomen, it is difficult to identify early.

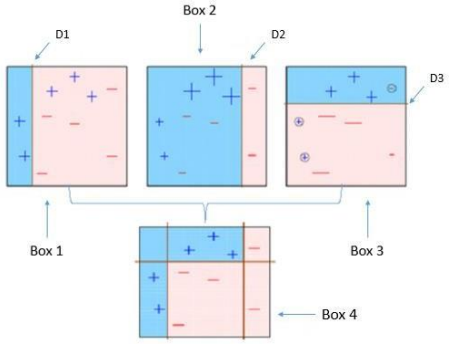


Fig. 4. Bagging & Boosting.

For this stage of the model generation, we'll utilize the same neural network configuration for all four situations. Layers, Perceptron, and layers are used to solve classification problems. We recognize that having a perceptron layer adds to the neural network being overfit. As a result, the perceptron layer is removed. Let's start with bagging techniques. The following equation demonstrates the principle of bagging, which is short for bootstrap aggregation: On a bootstrapped dataset, train several weak learners f b(x) and take the average to get the learning outcome. The term "bootstrap" refers to the process of producing different data samples from the original dataset at random (roll n-faces dice n times).

$$f(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

Following this logic, the Random Forest algorithm is naturally introduced, as decision trees are an excellent option for weak learners. Low bias and large variance are two features of a single decision tree. Bias remains after aggregating a group of trees, although variance decreases. By developing a large enough random forest, we could attain a constant bias that is as low as possible.

Let's move on to boosting now. The main principle behind boosting is to see if a poor learner can be made to improve by focusing on their weaknesses. This is accomplished by repeatedly employing the weak learning method to generate a series of hypotheses, each one focused on the cases that the prior hypotheses found problematic and misclassified.

$$f(x) = \sum_{t} \propto_t h_t(x)$$

E. Experimental Results

1) Testing analysis with similarity index: The performance of the trained neural network is subsequently evaluated utilizing an extensive testing analysis. The conventional way is to compare the neural network's outputs against previously unseen data, known as testing instances. The ROC curve is a well-known method for evaluating generalization performance. This is a visual aid for studying the discrimination capabilities of the classifier. One of the parameters acquired from this graph is the area under the curve (AUC). The closer the classifier is to 1 area under the curve, the better.

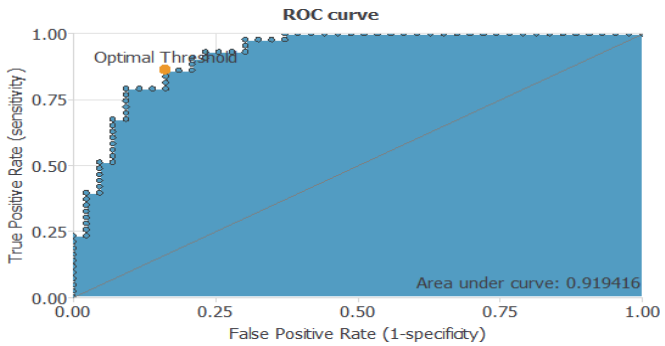


Fig. 5. ROC Curve for Control Samples & PDAC Stage 1 & 2.

a) Control Samples and PDAC Stage I and II: The AUC assumes a high value in this case: AUC = 0.919. The ideal threshold is calculated by identifying the point on the ROC curve (Fig. 5) that is closest to the upper left corner in Neural Designer. The ideal threshold is the one that corresponds to that point, and it has a value of 0.788 in this example. The confusion matrix and binary classification tests provide useful information regarding the performance of our predictive model. Both are shown below for their best choice threshold (Table I).

TABLE I. PREDICTIVE OF POSITIVE & NEGATIVE THRESHOLD STAGE 1 & STAGE 2

	Predictive Positive	**Predictive Negative**
Real Positive	34 (39.5%)	6 (7.0%)
Real Negative	9 (10.5%)	37 (43.0%)

Classification accuracy: 82.6 percent (Ratio of correctly classified samples), Error rate: 17.4 percent (Ratio of misclassified samples), Sensitivity: 79.1% (Proportion of true positive samples that are projected positive), and Specificity: 86.0 percent (Portion of real negative predicted negative). The classification accuracy is good (82.6%), indicating that the prediction is applicable to a broad number of scenarios.

b) Control Samples and PDAC Stage III and IV: The AUC takes a high value in this case: The ideal threshold is 0.587, and the AUC is 0.913 (Fig. 6 and Table II).

Classification accuracy: 88.6% (Ratio of correctly categorized samples), Error rate: 11.4 percent (Ratio of misclassified samples), Sensitivity: 92.4 percent (Percentage of genuine positive samples that are predicted positive), and Specificity: 81.3 percent (Portion of real negative predicted negative). The classification accuracy is good (88.6%), indicating that the forecast is applicable to a vast number of scenarios.

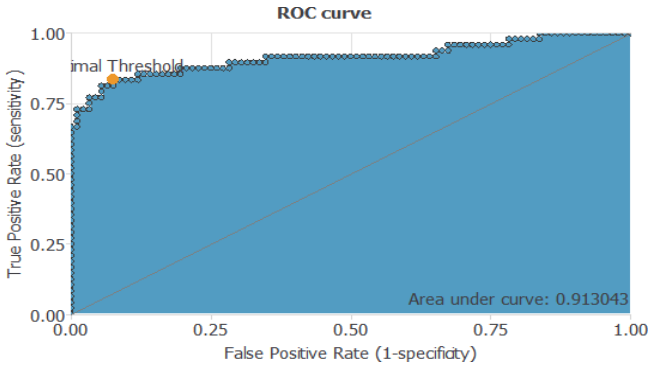


Fig. 6. ROC Curve for Control Samples & PDAC Stage 3 & 4.

TABLE II. PREDICTIVE OF POSITIVE & NEGATIVE THRESHOLD STAGE 3 & STAGE 4

	Predictive Positive	**Predictive Negative**
Real Positive	85 (60.7%)	9 (6.4%)
Real Negative	7 (5.0%)	39 (27.9%)

c) Difference between Benign Hepatobiliary diseases and PDAC Stage I and II: The AUC takes a high value in this case: The ideal threshold is 0.653, and the AUC is 0.920 (Fig. 7 & Table III).

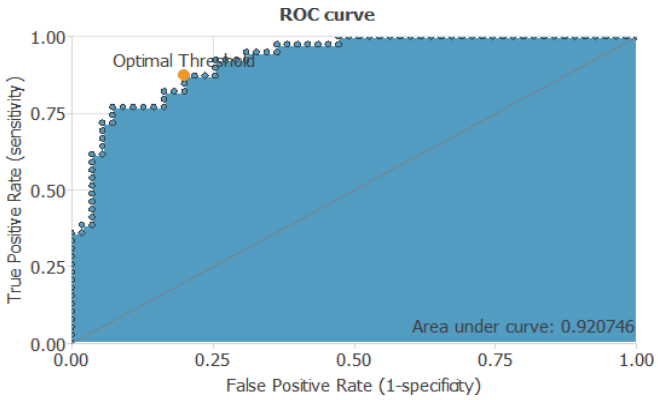


Fig. 7. ROC Curve for Benign Hepatobiliary Diseases & PDAC Stage 1 & 2.

TABLE III. PREDICTIVE OF POSITIVE & NEGATIVE THRESHOLD BENIGN HEPATOBILIARY STAGE 1 & STAGE 2

	Predictive Positive	**Predictive Negative**
Real Positive	44 (46.8%)	5 (5.3%)
Real Negative	11 (11.7%)	34 (36.2%)

Classification accuracy: 83.0% (Ratio of correctly classified samples), Error rate: 17.0% (Ratio of misclassified samples), Sensitivity: 80.0 percent (Proportion of true positive samples that are predicted positive), and Specificity: 87.2 percent (Portion of real negative predicted negative). The classification accuracy is good (83.0%), indicating that the forecast is applicable to a vast number of scenarios.

d) Difference between Benign Hepatobiliary Disease and Stage III and Stage IV: The AUC takes a high value in this case: The ideal threshold is 0.412, and the AUC is 0.848 & (Table IV).

TABLE IV. PREDICTIVE OF POSITIVE & NEGATIVE THRESHOLD BENIGN HEPATOBILIARY STAGE 3 & STAGE 4

	Predictive Positive	**Predictive Negative**
Real Positive	47 (52.8%)	11 (12.4%)
Real Negative	8 (9.0%)	23 (25.28%)

Classification accuracy: 78.7% (Ratio of correctly classified samples), Error rate: 21.3 percent (Ratio of misclassified samples), Sensitivity: 85.5 percent (Percentage of true positive samples that are projected positive), and Specificity: 67.6% (Portion of real negative predicted negative). The classification accuracy is good (78.7%), indicating that the forecast is appropriate in many circumstances. We'll show a table with some sensitivity and specificity cut-offs, just like in the paper. Table V, will look at the control samples vs. pancreatic cancer stages I and II, as well as stages III and IV:

TABLE V. CONTROL SAMPLES VS PANCREATIC CANCER STAGE 1 & 2

Sensitivity Cut-off	Specificity (Controls vs I, II)	Specificity (Controls vs III, IV)
0.8	0.86	0.875
0.85	0.791	0.854
0.9	0.744	0.833
0.95	0.512	0.771

Now Table VI will look at how benign samples compare to pancreatic cancer stages I and II, as well as stages III and IV:

TABLE VI. BENIGN SAMPLES VS PANCREATIC CANCER STAGE 3 & 4

Sensitivity Cut-off	Specificity (benign vs I, II)	Specificity (benign vs III, IV)
0.8	0.846	0.676
0.85	0.769	0.647
0.9	0.769	0.618
0.95	0.615	0.559

e) Deployment of the Model: The neural network can be preserved for future usage in the so-called model deployment mode once its generalization performance has been evaluated. Calculating outputs, which generates a set of outputs for each set of inputs given, is an interesting activity in the model deployment tool. The outputs, in turn, are determined by the parameter values. Fig. 8 then, for the benign tumour or PDAC stages III and IV diagnosis, will offer an example. LYVE1: 3.78856, REG1B: 121.787, TFF1: 752.305, diagnosis: 0.6895, age: 45, sex: F (1), plasma CA19- 9: 740.94, creatinine: 0.927814, LYVE1: 3.78856, REG1B:121.787, TFF1:752.305, diagnosis: 0.6895 That person's chance of pancreatic cancer (stages III or IV) would be high. Table VII and Table VIII shows the Model & Detailed Accuracy by Class.

Fig. 9 shows the Detailed Accuracy by Class and Fig. 10 Shows the Association between CCI (Correctly classified Instances) and ICUI (Incorrectly class Unknown Instances). Fig. 11 shows the association between CCI, ICCI, ICUI, and TNI.

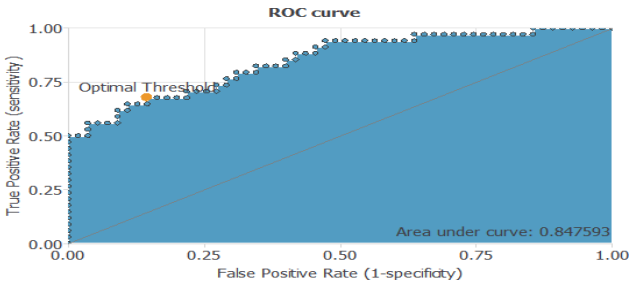


Fig. 8. ROC Curve for Benign Hepatobiliary diseases & PDAC Stage 3 & 4.

TABLE VII. MODEL ACCURACY

Algorithms	Instance of CCI	CCI	Instance of ICCI	ICCI	KS	MAE	RMSE	RAE	RRSE	ICUI	TNI
Navie Bayes	70	35.17%	129	64.8%	0.17	0.16	0.32	93.55	108.7	391	199
Nbtree	101	50.75%	98	49.24%	0.28	10.15	0.28	84.6551	96.71	391	199
Bagging	76	38.19%	123	61.80%	0	0.17	0.29	98.85	100.02	391	199
Adaboostml	82	41.20%	117	58.79%	0.16	0.168	0.29	93.10	97.69	391	199
Log Boosting	95	47.73%	104	52.26%	0.23	0.15	0.301	83.05	100.44	391	199

TABLE VIII. DETAILED ACCURACY BY CLASS

Algorithms	TP Rate	FP Rate	PRECISION	RECALL	F-MEASURE	ROCA
Navie Bayes	0.352	0.159	0.442	0.352	0.378	0.656
Nbtree	0.508	0.226	0.486	0.508	0.488	0.64
Bagging	0.382	0.382	0.146	0.382	0.211	0.481
Adaboostml	0.412	0.259	0.211	0.412	0.278	0.688
Log Boosting	0.477	0.245	0.46	0.477	0.454	0.691

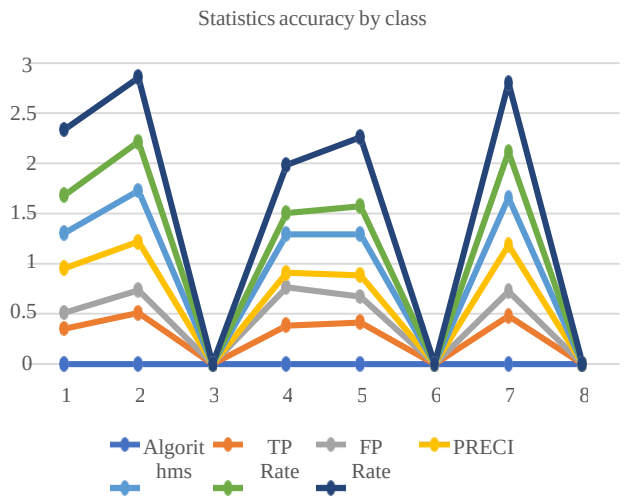


Fig. 9. Detailed Accuracy by Class.

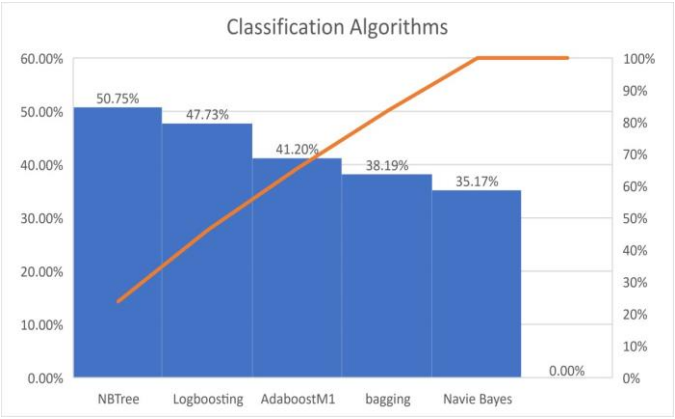


Fig. 10. Association between CCI and ICCI.

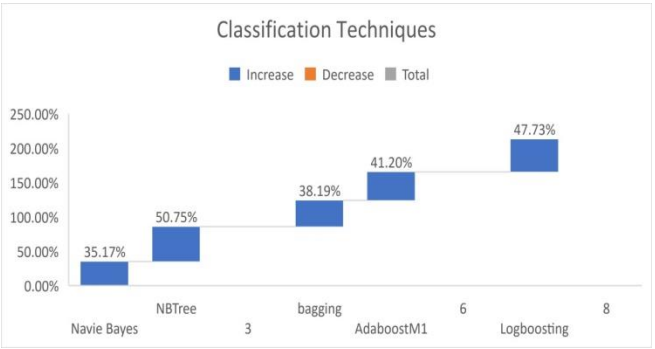


Fig. 11. Association between CCI, ICCI, ICUI, TNI.

IV. Conclusion

This study looks at how to use ensemble approaches in machine learning to analyse pancreatic tumours. Researchers are working to add features like active attention and in-line memory, which will allow folding neural networks to evaluate new elements that are significantly different from what they were trained on, and preliminary results show that the proposed approach can improve the classifier's performance for early detection of pancreatic cancer. This mirrors a mammalian visual system more closely, proposing a more intelligent artificial picture recognition categorization. Even though he collects invasive samples, he increases cancer diagnosis when combined with other urine indicators in a study. Previous research has found that a panel of three protein biomarkers (LYVE1, REG1A, and TFF1) found in urine can help detect significant PDAC. We improved this panel in this study by replacing REG1A with REG1B. Finally, we will analyse four significant biomarkers that are found in urine: creatinine, LYVE1, REG1B, and TFF1. Creatinine is a protein that is commonly utilised as a kidney function indicator. Lymphatic vessel endothelial hyaluronan receptor 1 (YVLE1) is a protein that may help tumours spread. REG1B is a protein that has been linked to pancreatic regeneration, while TFF1 is trefoil factor 1, which has been linked to urinary tract regeneration and repair. This regularisation of the form's continuity allows for the smoothness of pancreatic segmentation. The preliminary result reflects the state of the art in pancreatic cancer prediction and reaches a high level of precision. However, further study is needed to detect early pancreatic cancer, because COVID-19 infection-induced pancreatic damage has gotten minimal attention. Moving further, we must compare the mood analysis of Twitter API with COVID-19 examples for pancreatic cancer detection and apply advanced innovation algorithms to existing Hadoop ecosystem work using deep learning and learning paradigms in the goal of early pancreatic cancer detection. As additional samples from various central institutions are collected and the best-performing classification model is established, preoperative diagnosis and staging from a computer using samples will be of substantial therapeutic benefit in the future.

References

[1] H. Matsubayashi, H. Ishiwatari, K. Sasaki, K. Uesaka, and H. Ono,``Detecting early pancreatic cancer: Current problems and future prospects," Gut Liver, vol. 14, no. 1, pp. 3036, Jan. 2020.

[2] Z.-Y. Wang, X.-Q. Ding, H. Zhu, R.-X. Wang, X.-R. Pan, and J.-H. Tong,``KRAS mutant allele fraction in circulating cell-free DNA correlates with clinical stage in pancreatic cancer patients," Frontiers Oncol., vol. 9,p. 1295, Nov. 2019.

[3] Morris, J. P.; Cano, D. A.; Sekine, S.; Wang, S. C.; Hebrok, M. beta-catenin blocks Kras-dependent reprogramming of acini into pancreatic cancer precursor lesions in mice. J. Clin. Invest. 2010, 120, 508−520.

[4] Shamsaldin, A. S., Rashid, T. A., Al-Rashid Agha, R. A., Al-Salihi, N. K., &Mohammadi, M. (2019). Donkey and smuggler optimization algorithm: A collaborative working approach to path finding. Journal of Computational Design and Engineering, 6(4), 562-583.

[5] S. Liu, X. Yuan, R. Hu, S. Liang, S. Feng, Y. Ai, and Y. Zhang, ``Automatic pancreas segmentation via coarse location and ensemble learning," IEEE Access, vol. 8, pp. 29062914, 2020.

[6] Suram, A.; Kaplunov, J.; Patel, P. l.; Ruan, H.; Cerutti, A.; Boccardi, V.; Fumagalli, M.; Di Micco, R.; Mirani, N.; Gurung, R. L.; Hande, M. P.; d'Adda di Fagagna, F.; Herbig, U. Gurung. Oncogeneinduced telomere dysfunction enforces cellular senescence in human cancer precursor lesions. EMBO J. 2012, 31, 2839−2851.

[7] Glicksberg BS, Miotto R, Johnson KW, Shameer K, Li L, Chen R, Dudley JT (2018) Automated disease cohort selection using word embeddings from Electronic Health Records. Pac SympBiocomput.

[8] Miotto R, Li L, Dudley JT (2016) Deep learning to predict patient future diseases from the electronic health records. European Conference on Information Retrieval.

[9] Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, Albuquerque K, Gu X (2017) Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. Institute of Physics and Engineering in Medicine Physics in Medicine & Biology 62.

[10] Fave, X. et al. Using pretreatment radiomics and delta-radiomics features to predict nonsmall cell lung cancer patient outcomes. Int. J. Radiat. Oncol. Biol. Phys. 7, 588 (2017).

[11] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, . J. Van Arnam, Spatialorganization and molecular correlation of tumor-infiltrating lymphocytesusing deep learning on pathology images, Cell Rep. 23 (1) (2018) 181–193.

[12] Y.H. Chang, G. Thibault, O. Madin, V. Azimi, C. Meyers, B. Johnson, . J.W.Gray, Deep learning- based nucleus classification in pancreas histologicalimages, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017,pp.672–675.

Developing and Validating Instrument for Data Integration Governance Framework

Noor Hasliza Mohd Hassan, Kamsuriah Ahmad, Hasimi Salehuddin
Center for Software Technology and Management, Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

***Abstract*—Data integration is one of the important subfields in data management. It allows users to access the same data from multiple sources without redundancy and preserving its integrity. Data Integration Governance Framework (DIGF) is being developed to guide the implementation of data integration. It functions as a reference and guideline for working level in data integration implementation. Hence the instrument used to validate the DIGF needs to be developed and validated for its accuracy, applicability, and suitability of use. The instrument comprises items structured as a questionnaire. This study proposes Lawshe's technique to construe the content validity of the instrument. This technique involved the arithmetic of the Content Validity Ratio (CVR) to validate items in the questionnaire, which developed based on the factors identified for Data Integration Governance Framework. Each item in the questionnaire that validated based on the minimum CVR value of 0.75 endorsed as the final instrument of Data Integration Governance Framework to be used in Delphi Technique Evaluation.**

Keywords*—Content validity; instrument development; data integration governance; Lawshe's technique

I. Introduction

Data integration plays an important role to provide cleaned, integrated, and secured data for decision making and operation purposes in public sector [1], [2], [3]. Public sector as the biggest owner of data, needs an efficient data integration governance to support the digitalization plan [4]. An efficient governance should incorporate all aspects influencing data integration governance in public sector. Previous studies show that, focusing only on aspect of technology will prompt to failure in data integration governance [5], [6], [7].

Thus, to solve the issue, this study identified dimensions and factors influencing data integration governance in public sector using literature review, theories adoption and interview method. The factors and dimensions identified later being constructed into the public sector Data Integration Governance Framework (DIGF). DIGF that has been developed needs to be validated to ensure it suits the practicality and requirement of the public sector data integration initiative.

This study uses Delphi Technique to validate the framework using the questionnaire with Likert Scale to measure the validity of the framework. However, the Delphi Technique process should be preceded with an instrument validation [8], [9]. Hence, this study will focus on implementing content validation process to validate the instrument will be used for Delphi Technique. Content Validation Ratio (CVR) and Index of Content Validity (CVI) are identified as the measurement method for this process.

This paper will be segmented into three major parts, which are firstly, the description of DIGF, comprising the dimensions and factors explanation; secondly, the methods used with the questionnaire summary; and lastly, the results and discussion on the data analysis.

II. Related Work

A. DIGF

Three dimensions that have been identified in this study are people, process, and technology. Meanwhile, the factors listed are culture, clarity of roles and responsibility, and communication under the people dimension; law and regulation, and policy under process dimension and for the technology dimensions, factor is summed up as tools and technology. The relationship between all the three dimensions and six factors is being employed as the foundation for the DIGF development. DIGF development involves literature review of previous study, theories adoption and interview with the experts, to simulate and correlate the dimensions and factors influencing data integration governance in public sector. The description of the dimensions and factors are given in Table I.

Based on the description and connection between the dimensions and factors above, this study has come out with a framework of data integration governance in public sector. The framework developed as per in Fig. 1.

Public sector DIGF that has been erected is a strategic basic framework as the dimensions and factors are connected and described generally. It could be a reference and adapted to any kind of organization including private sector in governing their data integration initiative.

B. Development of Questionnaire

Heeding to a rigid protocol suggested by [16] and [17], there are four processes and six supporting steps of developing questionnaire in content validation process. Notwithstanding, this study has come out with four processes and eleven supporting steps of developing questionnaire in content validation process. Fig. 2 explains the process adapted by this study in questionnaire development.

TABLE I. DESCRIPTION OF DIMENSIONS AND FACTORS OF DIGF

Dimension	Factor	Description
People People refers to the entity that perform the activities using the tools and technology provided according to the objective ad principle set up. [10]	Culture	Culture involves knowledge, beliefs, habits, capability, and norms in an organization that influence the individual's and organization's goals. [11]
	Clarity of roles and responsibility	Roles and responsibility described the contribution of the personnel towards the activities in the organization commensurate to their expertise and qualification. Clarity of roles and responsibility gives impact to facilitate the governance of any initiative. [10]
	Communication	Communication relates to human's behavior. It also applied to other entities such as the hardware and software. Communication basically connects all the dimensions and factors together.[12],[13]
Process Process is a set of related activities with input, value add, and procedures which produces specific output. Process automated by the technology and facilitated by people.[14]	Law and regulation	Law and regulation cover the law (act) and official orders issued by the government or the authorities to control or govern the implementation of activities and human behavior.
	Policy	Policy is a simple and comprehensive mandatory formal statement that outline the rules and commands for an organization in performing any activities.
Technology Technology refers to the tools and techniques used by people in implementing any activities. Technology creates innovative human resources and automated the processes.[14], [15]	Tools and technology	Tools and technology is the factor that support and facilitate the process and people's task. [10], [14], [15]

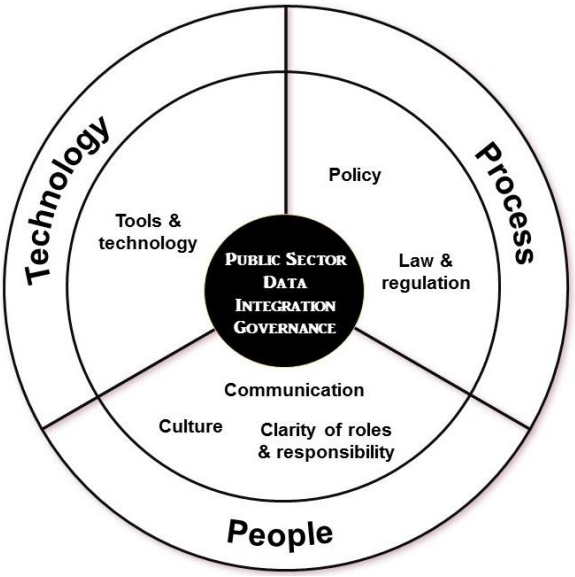


Fig. 1. Data Integration Governance Framework.

The questionnaire booklet is segmented into four components, which are (1) panel information, (2) items on factors influencing data integration governance, (3) definition of data integration and data integration governance, and (4) Data Integration Governance Framework (DIGF). Component (1) used for collecting panel's information such as job designation, place of work, year of experience and contact information. Meanwhile, component (2) consists of 94 specific items and 6 generic items on factors embedded in DIGF. Component (3) includes the definition of data integration and data integration governance that needs to be validated by the experts and component (4) covers the explanation of DIGF.

Specific items refer to the individual items for each factor. Meanwhile, the generic items represent the whole factor in general. Generic items is important to be developed as it gives opportunity to the experts to evaluate the factors in general [18]. The content of the items listed in the questionnaire for DIGF validation is summarized in Table II. Items were developed based on literature review, whereby discussed by previous studies.

Process 1: Planning and Strategizing
Step 1: Define type of method and questionnaire Step 2: Clarify administrative process
Process 2: Defining Content
Step 3: Provide conceptual definition of dimensions and factors Step 4: Develop items for factors Step 5: Define measurement skills Step 6: Identify the experts
Process 3: Designing Questionnaire
Step 7: Design and develop the questionnaire
Process 4: Validating Questionnaire
Step 8: Conduct content validation process Step 9: Calculate CVR value Step 10: Calculate CVI value Step 11: Analyse the results

Fig. 2. Questionnaire Development Process.

TABLE II. HIGHLIGHT OF ITEMS IN QUESTIONNAIRE

Factors	Item Code	Highlight of items
Dimension A: People		
Factor A1 – Culture	Specific: A1-1 till A1-16 Generic: 4	• Culture includes the aspects of organization culture and individual culture as the member of the organization. • Good organization culture ensures the alignment between corporate strategy and IT strategy. • Data sharing culture through data integration initiative plays an important role in data in public sector. • To successfully implementing data integration initiative, organization culture and individual work culture must be aligned and understood well. • Ethics must be incorporated as an organization culture in data management area especially in data integration.
Factor A2 – Clarity of roles and responsibilities	Specific: A2-1 till A2-15 Generic: 7	• Clarity of roles and responsibilities could ensure the member of organization assimilate their job scopes and task in data integration governance. • Clear power and job scope distribution determine the accountability and responsibility of the member of organization. • Efficient leadership spearheading an effective data integration governance in organization.
Factor A3 – Communication	Specific: A3-1 till A3-15 Generic: 10	• Communication is an enabler to ensure other factors could be adapted efficiently. • Clear, structured, and effective communication will help the member of organization to comprehend the objectives, terms of reference and planning of data integration initiative in organization • The benefits of data integration should be communicated to the member of organization for them to support the implementation. • The usage of data standard and standard term in integration team will assist in data integration implementation. • Organization needs to provide effective communication channels to facilitate data integration governance. • Organization needs to provide an effective change management plan and execution to facilitate data integration governance.
Dimension B: Process		
Factor B1 – Law and regulation	Specific: B1-1 till B1-18 Generic: 13	• Law and regulation include establishing act to protect and guide data integration governance. • Former acts regarding data integration and data sharing should be updated and aligned. • There should be an act enforced to protect the data security, privacy, and confidentiality. • Alignment between federal, state, and local council's law and regulation should be established to support data integration initiate in public sector.
Factor B2 - Policy	Specific: B2-1 till B2-16 Generic: 16	• Policy in organization or public sector itself helps to determine the direction, guideline, and rules in data integration implementation in public sector. • Establishment of clear and systematic policy will lead into good data integration governance and efficient implementation. • Policy alignment between federal, state, and local council should be established to support data integration initiate in public sector.
Dimension C: Technology		
Factor C1 – Tools and technology	Specific: C1-1 till C1-14 Generic: 19	• Choosing the right technology is crucial to assure compatibility, maintainability, reliability, and security of data integration initiative. • Choosing the right technology also will provide high quality data through well equipped function such as data cleansing, data profiling, data stewardship, and others. • Tools and technology selection must be aligned and complied to the law and regulation, policy, cultural, and organization corporate and IT strategy.

III. MATERIAL AND METHOD

Delphi Technique has been identified as the validation method for DIGF. Delphi technique involves getting consensus from the experts to validate research output in an iteration process [19], [20]. However, the questionnaire that will be used as an instrument in Delphi Technique need to be validated in a pilot study to ensure that the perceived construct are clear, valid and manifest its contents [17], [18]. The process is also known as content validation process.

There are many content validation methods available such as psychometric analysis using Rasch Model [8] and modified kappa statistic [21]. However, this study recognized CVR and CVI as the methods to validate the Delphi Technique instrument as it involve experts' evaluation and commonly used for content validation for Delphi Technique [17], [18].

A. Selection of Experts

Experts' selection would be the most crucial and initial part of content validation process. Among the criteria of experts' selection are; (1) technical knowledge and experience in the research area, (2) willingness to participate, (3) having ample time to involve in the process and (4) possessing good communication skill [22], [23]. In this study, experts were selected based on their experience and knowledge in data integration area, research process and Delphi Technique. The numbers of experts selected is normally based on the research scope, resources available, (which include time and cost) and research objectives [24], [25]. Nonetheless, there is no definite mechanism to determine the right numbers of experts involved in content validation process and Delphi Technique for every different research [26].

Eight experts were identified for content validation process based on the criteria and requirement set up for this study as per in Table III.

TABLE III. EXPERTS' CRITERIA

Requirement aspect	Criteria
Practicality aspect of content	10 years or more experience in data management and/or data integration area
Methodology and academic content	10 years or more experience in research area
Academic language and methodology	5 years or more experience in research using Delphi Technique and published article related to Delphi Technique or CVR

B. Content Validity Ratio (CVR)

Content Validity Ratio (CVR) was introduced by Lawshe in 1975 [27]. This method has been widely used in many research domains including computer science and engineering. According to Google Scholar, up to January 2022, Lawshe's CVR technique has been referred and cited for 7,149 times in various research publication. Meanwhile, review in Scopus Database identified 19 research on computer science and engineering from year 2016 until 2020 using CVR method to validate their content including three research that validated content of instrument for Delphi Technique using CVR [17], [28], [29].

CVR uses Likert Scale with three indicators, which are, "1-not necessary", "2-useful (but not essential)" and "3-essential". Likert Scale with 3 indicators is being used to simplify and provide an objective evaluation process by the experts [17], [18]. Comments section is provided for the experts to express their opinion and suggestion of improvement on the items.

CVR calculation and analysis assess the experts' agreement on the listed items in the questionnaire using formula introduced by Lawshe as below:

$$CVR = \frac{n_c - N/2}{N/2} = \frac{2n_c - 1}{N}$$

Where n_e is numbers of experts picked scale 2 and 3, and N equal to total numbers of experts.

Precondition for n_e is based on the suggestion by [17], [18], and [30], as they concluded that indicators "2-useful (but not essential)" and "3-essential" refer to positive feedback from the experts which conduce to acceptance of the items.

The conditions for the formula by Lawshe are as below:

1) If all experts answer "3-essential", CVR value equal to 1.00.

2) If more than half experts (>50%) but less than 100%, answer "3-essential", CVR value is between 0 and 0.99.

3) If less than half (>50%) experts answer "3-essential", CVR value will be a negative value.

However, for this study, as suggested by [17] and [18], indicator "2-useful (but not essential)" also accepted, as both, indicators 2 and 3 reflect positive acceptance and relevancy to the study. This study also follows recommendation of [27] on minimum CVR value for items' acceptance based on numbers of experts participated as in Table IV.

Considering the numbers of experts participated in this study is eight (refer Table V), the accepted minimum CVR value is 0.75 for each item. All items that obtain minimum CVR value of 0.75 will proceed to the Delphi Technique process.

TABLE IV. MINIMUM CVR VALUE

No. of experts	Minimum CVR value
5	.99
6	.99
7	.99
8	.75
9	.78
10	.62
11	.59
12	.56
13	.54
14	.51
15	.49
20	.42
25	.37
30	.33
35	.31
40	.29

C. Index of Content Validity (CVI)

CVI is being used to evaluate the whole instrument either it measures the right and relevant items that should be measured or otherwise [31]. According to [32], as content validation process is very important to endorse the instrument of the study, it need to be done in a systematic arrangement with strong justification and credible proof. In this study, the CVI calculation used is adapted from [27] and [33] which CVI is equal to mean of CVR. The calculation of CVI where 't' is the accepted items is demonstrated as below.

$$Mean_{CVR} = \frac{Total\ CVR}{No.of\ accepted\ items} = CVI$$

$$\mu_{CVR} = \frac{\sum CVR_t}{\sum t} = CVI$$

According to [33], the closer value of CVI to 0.99 the higher value of content validity we should get. This means, the level of acceptance for the whole instrument will be higher too.

IV. RESULTS AND DISCUSSION

A. Selection of Experts

Experts were selected based on criteria determined in Table III. Eight experts from public sector and academics had been selected and agreed to participate in this study. The list of experts as per stated is in Table V.

B. Questionnaire Distribution

Content validation process was done within two weeks. Invitation was done through email and followed up by telephone call. Experts who agreed to participate will receive official invitation from the faculty and the questionnaire then distributed through email. Further explanation was done using email, telephone call and "WhatsApp" accordingly.

TABLE V. LIST OF EXPERTS

No.	Designation	Organization
1	Principle Assistant Director	Malaysian Administrative Modernisation and Management Planning Unit (MAMPU)
2	Principle Assistant Director	Attorney General Chambers of Malaysia
3	Senior Assistant Director	MAMPU
4	Associate Professor	Universiti Kebangsaan Malaysia
5	Senior Assistant Director	MAMPU
6	ICT Consultant	MAMPU
7	Senior lecturer	Universiti Putra Malaysia
8	Senior lecturer	Universiti Malaya

C. Analysis of Questionnaires

The minimum value of CVR accepted as mentioned above is based on numbers of experts involved. For this study, the minimum value accepted is 0.75 as we have eight experts on board. By this means, both specific and generic items with CVR value equal to 0.75 and above will be brought to the first round of Delphi Technique process for DIGF validation.

Content validation analysis based on Lawshe's technique is presented as below.

1) All item obtained CVR value of 1.00 except seven items obtained 0.75 CVR value. The seven items with 0.75 CVR value are item 1.A1-3, 1.A1-4 and item 1.A1-5 for Factor A1- Culture, item 5.A2-2 and 5.A2-6 for Factor A2- Clarity of roles and responsibility, item 11.B1-5 on Factor B1- Law and regulation, and the last one, item 18.C1-10 and 18.C1-13 on Factor C1- Tools and technology.

2) Item 1.A1-3 described that organization culture should not be a limitation in data integration initiative. Item 1.A1-4 stated that organization culture should be considered to design a data integration initiative in an organization. Meanwhile, item 1.A1-5 suggested that organization must ensure there would be no conflict of culture while adopting data integration in organization. There is one answer with indicator "1=not necessary" for these three items. However, the experts did not leave any comment on the items.

3) Item 5.A2-2 described that clarity of roles and responsibility will support member of organization to perform their task at their best capability and skill. Meanwhile, item 5.A2-6 explained that clarity of roles and responsibility will balance and incorporate technical and management aspects in data integration governance. For each item, there is one expert answered "1=not necessary". However, no comments were provided by the experts on these items.

4) Item11.B1-5 explained that managing law and regulation factor is important in data integration governance so that it would not be an obstacle in new technology adoption and utilization. An expert picked "1=not necessary" for this item with no comment provided.

5) Item 18.C1-10 and item 18.C1-13 received one "1=not necessary" each from one expert. Item 18.C1-10 stated that technology need to be accommodated with human resource capability in the organization. Expert's comment on item 18.C1-10, "human resource needs to adapt with technology and not other way around". Item 18.C1-13 describe those tools and technology adopted must be free from vendor lock-in. No comments received for item 18.C1-13.

6) As all items obtained CVR value of 0.75 and above, all items are accepted and bring forward to the first round of Delphi Technique.

7) For generic item, all six items earned CVR value 1.00. This demonstrates that all experts agreed upon the importance of every factor equipped in DGIF.

8) All factors earned CVI more than 0.95 and the overall CVI for the questionnaire is 0.98. This concludes that overall questionnaire is measuring the right things for DIGF and validated by the experts.

Summary of CVR and CVI calculation for 94 items included in the questionnaire is presented in Table VI.

TABLE VI. SUMMARY OF CVR AND CVI ANALYSIS

Dimension and factor	CVR value specific item	CVR value generic item	CVI
Dimension A – People			
Factor A1 – Culture	All item = 1 (Except item 1.A1-3, 1.A1-4 and 1.A1-5 = 0.75)	Item 4 = 1	0.95
Factor A2 – Clarity of roles and responsibility	All item = 1 (Except item 5.A2-2 and 5.A2-6 = 0.75)	Item 7 = 1	0.97
Factor A3 - Communication	All item = 1	Item 10 = 1	1.00
Dimension B – Process			
Factor B1 – Law and regulation	All item = 1 (Except item 11.B1-5 = 0.75)	Item 13 = 1	0.99
Factor B2 - Policy	All item = 1	Item 16 = 1	1.00
Dimension C – Technology			
Factor C1 – Tools and technology	All item = 1 (Except item 18.C1-10 and 18.C1-13 = 0.75)	Item 19 = 1	0.96
Overall CVI			**0.98**

V. CONCLUSION

From the analysis executed, all 94 specific items and six generic items developed in the questionnaire are accepted by the experts. This indicate that items attached to the six factors included in DIGF have been validated through the content validation process using CVR and CVI calculation based on Lawshe's Technique. In conclusion, this questionnaire has been validated by the experts through content validation process and now ready to be used in Delphi Technique process to validate the DIGF. The validated DIGF will then be adopted in ensuring the successful implementation of data integration initiatives.

ACKNOWLEDGMENT

This study is sponsored by the Geran Galakan Penyelidikan (GGP) UKM, (Grant No. GGP-2019-024), and supported by Centre for Software Technology and Management (SOFTAM) of Faculty of Information Science and Technology, National University of Malaysia (UKM) and Public Service Department (PSD) of Prime Minister's Department.

REFERENCES

[1] T. Yang and T. A. Maxwell, "Information-sharing in public organizations : A literature review of interpersonal , intra-organizational and inter-organizational success factors," Gov. Inf. Q., vol. 28, no. 2, pp. 164–175, 2011.

[2] L. Zheng, S. Dawes, and T. A. Pardo, "Leadership Behaviors in Cross-boundary Information Sharing and Integration : Comparing the US and China," in ICEGOV2009, 2009, pp. 43–50.

[3] R. Omar, T. Ramayah, M. Lo, T. Y. Sang, and R. Siron, "Information sharing , information quality and usage of information technology (IT) tools in Malaysian organizations," African J. Bus. Manag., vol. 4, no. 12, pp. 2486–2499, 2014.

[4] R. Munne, "Big Data in the Public Sector," in New Horizons for a Data-Driven Economy, J. María, C. Edward, and W. Wahlster, Eds. 2016, pp. 195–208.

[5] A. . Amadi-Echendu and J. . Amadi-Echendu, "A Study on Data and Information Integration for Conveyancing, Cadastre and Land Registry Automation," in 2016 Proceedings of PICMET '16: Technology Management for Social Innovation, 2016, pp. 804–814.

[6] S. Eom and J. H. Kim, "Information Sharing Success in Korean Metropolitan Governments : Combining Multi-level Factors with Fuzzy-set Analysis," in Proceedings of dg.o conference, Staten Island, New York USA, June 2017 (dg.o'17), 2017, pp. 279–288.

[7] M. I. Manda, "Towards 'Smart Governance' Through a Multidisciplinary Approach to E-government Integration, Interoperability and Information Sharing : A Case of the LMIP Project in South Africa," in International Federation for Information Processing 2017, 2017, pp. 36–44.

[8] S. Sanz-Martos, I. M. López-Medina, C. Álvarez-García, and C. Álvarez-Nieto, "Sexuality and contraceptive knowledge in university students: Instrument development and psychometric analysis using item response theory," Reprod. Health, vol. 16, no. 1, pp. 1–11, 2019.

[9] B. R. Lewis, G. F. Templeton, and T. A. Byrd, "A methodology for construct development in MIS research," Eur. J. Inf. Syst., vol. 14, no. 4, pp. 388–400, 2005.

[10] I. DAMA, DAMA-DMBOK2 Data Management Body of Knowledge, 2nd Editon. Technics Publications, 2017.

[11] B. A. Blumenthal, "A new definition of culture," Am. Anthr., pp. 571–586, 1940.

[12] D. S. Sayogo, J. R. Gil-garcia, F. A. Cronemberger, and B. Widagdo, "The Mediating Role of Trust for Inter-Organizational Information Sharing (IIS) Success in the Public Sector," in 18th Annual International Conference on Digital Government Research, 2017, pp. 426–435.

[13] N. H. Mohd Hassan and K. Ahmad, "A Review on Key Factors of Data Integration Implementation in Public Sector," no. July, pp. 9–10, 2019.

[14] M. Prodan, A. Prodan, and A. A. Purcarea, "Three New Dimensions to People , Process , Technology Improvement Model," Adv. Intell. Syst. Comput., pp. 481–490, 2015.

[15] H. Leavitt and B. Bass, "Organizational Psychology," Annu. Rev. Psychol., vol. 15, no. 1, pp. 371–398, 1964.

[16] A. Ab Aziz, Z. M. Yusof, and U. A. Mokhtar, "Electronic Document and Records Management System Adoption : Instrument Development Protocol and Content Validation using Content Validation Ratio Electronic Document and Records Management System (EDRMS) Adoption in Public Sector – Instrument ' s Conte," J. Phys. Conf. Ser., vol. 1196, pp. 1–9, 2019.

[17] W. Azlin, Z. Wan, M. Mukhtar, and Y. Yahya, "Developing and Validating an Instrument for Social Content Management," vol. 10, no. 1, pp. 239–245, 2020.

[18] N. Ali, A. Tretiakov, and D. Whiddett, "A Content Validity Study for a Knowledge Management System Success Model in Healthcare," JITTA J. Inf. Technol. Theory Appl., vol. 15, no. 2, pp. 21–36, 2014.

[19] S. Siraj and A. Ali, "Principals Projections on the Malaysian Secondary School Future Curriculum," Int. Educ. Stud., vol. 1, no. 4, pp. 61–78, 2008.

[20] M. R. Mohd Jamil and N. Mat Noh, "Pengenalan Asas Kaedah Delphi," in Kepelbagaian Metodologi dalam Penyelidikan Reka Bentuk dan Pembangunan, 1st ed., I. Noh, Ed. Qaisar Prestige Resources, 2020, pp. 45–84.

[21] V. Zamanzadeh, A. Ghahramanian, M. Rassouli, A. Abbaszadeh, H. Alavi-, and A.-R. Nikanfar, "Design and Implementation Content Validity Study : Development of an instrument for measuring Patient-Centered Communication," J. Caring Sci., vol. 4, no. 5, pp. 165–178, 2015.

[22] N. Nordin, B. M. Deros, D. A. Wahab, and M. N. A. Rahman, "Validation of lean manufacturing implementation framework using delphi technique," J. Teknol. (Sciences Eng., vol. 59, no. 2, pp. 1–6, 2012.

[23] G. J. Skulmoski, F. T. Hartman, and J. Krahn, "The Delphi Method for Graduate Research," J. Inf. Technol. Educ., vol. 6, no. 1, pp. 1–21, 2007.

[24] A. L. Delbecq, H. Van de Ven, and David H. Gustafson, "Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes," J. Appl. Behav. Sci., vol. 12, p. 581, 1976.

[25] S. J. van Zolingen and C. A. Klaassen, "Selection processes in a Delphi study about key qualifications in Senior Secondary Vocational Education," Technol. Forecast. Soc. Change, vol. 70, no. 4, pp. 317–340, 2003.

[26] P. L. Williams and C. Webb, "The Delphi technique: a methodological discussion," J. Adv. Nurs., vol. 19, no. 1, pp. 180–186, 1994.

[27] C. . Lawshe, "A Quantitative Approach To Content Validity," Pers. Psychol., vol. 28, no. 4, pp. 563–575, 1975.

[28] Y. S. Hwang et al., "Current issues and areas for improvement in the Korean Dental Hygienist National Licensing Examination: an expert Delphi survey among dental hygienists," J. Educ. Eval. Health Prof., vol. 14, p. 21, 2017.

[29] N. M. Noh, S. Siraj, S. H. Halili, M. R. M. Jamil, and Z. Husin, "Application of fuzzy delphi method as a vital element in technology as a tool in design thinking based learning," Asia Pacific J. Educ. Educ., vol. 34, pp. 129–151, 2019.

[30] V. K. Shrotryia and U. Dhanda, "Content Validity of Assessment Instrument for Employee Engagement," SAGE Open, vol. 9, no. 1, 2019.

[31] M. R. Lynn, "Determination and quantification of content validity," Nurs. Res., vol. 35, no. 6, pp. 382–385, 1986.

[32] Y. Muhamad Saiful Bahri, "ABC of Content Validation and Content Validity Index Calculation," Educ. Resour., vol. 11, no. 2, pp. 49–54, 2019.

[33] T. Allahyari, N. H. Rangi, Y. Khosravi, and F. Zayeri, "Development and Evaluation of a New Questionnaire for Rating of Cognitive Failures at Work," Int. J. Occup. Hyg., vol. 3, no. 1, pp. 6–11, 2011.

The Method of Braille Embossed Dots Segmentation for Braille Document Images Produced on Reusable Paper

Sasin Tiendee[1], Charay Lerdsudwichai*[2]
Department of Computer Engineering
Faculty of Engineering, Kasetsart University
Bangkok, Thailand

Somying Thainimit[3]
Department of Electrical Engineering
Faculty of Engineering, Kasetsart University
Bangkok, Thailand

Chanjira Sinthanayothin[4]
National Electronics and Computer Technology Center
National Science and Technology Development Agency
Pathum Thani, Thailand

***Abstract*—Braille is the language of communication for blind and visually impaired people. Braille characters are embossed at points to convey the meaning. Typically, Braille documents can be produced on plain paper. Braille documents can be created on reusable paper, also known as a third-page paper; this reduces the paper cost, allowing more available documents to stimulate learning for blind or visually impaired persons. This research presents a method of Braille embossed dots segmentation for Braille document images produced on reusable paper to support the availability of cheaper learning material. Initially, Braille documents were imported with a calibrated scanner, Braille document image layer separation was then performed. Followed by edge removal, Braille embossed dot recovery, noise removal, and specify the embossed Braille point. This research was conducted by using four scanners, which scanned Braille documents images under four different lighting conditions. For each lighting condition, the Braille document image area was cropped to the desired size, considering the possible event conditions. They were used to create over 200,000 Braille cells, with over 12 billion patterns. When calculating the average performance under all lighting conditions, the values were Precision 1.0000, Recall 0.7817, Accuracy 0.8545, and F-Measure 0.8756. By effectively using Braille embossed dots segmentation, the process of Braille document recognition will also be efficient.**

Keywords*—Braille; embossed dots; document images; reusable paper; segmentation; recognition; blind; visually impaired

I. Introduction

Louis Braille invented Braille so allowing blind or visually impaired persons to communicate using written communication; subsequently, there is a requirement they become proficient in writing and reading Braille. In Braille, one cell of Braille has six dots that represents a meaning. To create Braille documents is a writing pad (slate) and a sharp tip (stylus) being portable and cheap. Other devices can create Braille documents, such as a Braille typewriter and a Braille printer, but these devices are expensive and require expensive specialized paper. It is common practice that blind, or the visually impaired people produce Braille documents on reusable paper, known as the third page. These documents contain Braille embossed dots, characters, tables, and pictures known as patterns. It reduces the cost of purchasing Braille paper and is a cost-effective use of natural resources.

Braille documents created with reusable paper are everywhere. This research aims to accurately extract the Braille embossed dots [1–6] on those Braille document. Those Braille embossed dots are used for Braille recognition [7–12] and converted to characters. In the end, these characters will be used to make typical books.

Scope and limitations: This research created reusable papers with characters printed by using LaserJet and Inkjet printers on an A4 80 GSM thick. Braille documents were created on the reusable papers by using a portable Braille device and scanned with a flatbed scanner at 300 DPI resolution.

Contribution: (1) To develop a method for Braille embossed dots segmentation for Braille documents produced on reusable papers. (2) To reduce the complexity problems and cost of purchasing paper to create Braille documents. (3) To support improved communication channels between people.

The paper was organized: Section II summarizes the relevant research and describes the new approaches to this research. Section III explains the proposed method. The dataset, experimental design, evaluation, and discussion of the results are described in Section IV. Section V summarizes the results and discusses them. The final section, Section VI, outlines our future work.

II. Related Work

The evolution of Braille document image processing is shown in Fig. 1. It is divided into two groups: (1) Group 1 document image processing for typical documents. (2) Group 2 document image processing of Braille documents created on plain paper for the visually impaired where the Braille characters have embossed dots.

Group 1 can be divided into 2 subgroups: (1) Subgroup 1 is documents created on plain paper without an overlaid pattern. There are relevant research topics such as text/non-text classification in online handwritten notes [16], the 2D chemical structures recognition in document images [17], detecting math

* Corresponding Author

equations in scientific document images [18], the Arabic word recognition of historical documents images [19], the Vietnamese character recognition for verifying ID card [20], document zoning for document layout analysis [21], analysis of the structure of the musical document image [22], bibliographic reference extraction [23], extracting text and figure from document images [24, 25], document localization in natural scene images [26], and table detection and segmentation in document images [27]. (2) Subgroup 2 is created on plain paper and overlapped patterns. There are related research topics such as image restoration and segmentation of historical document images caused by ink bleeding [28–30], glare detection on captured document images [31], shadow removal on captured document images [32], and text segmentation from a highlighted area with colors [33].

Group 2 has one subgroup: Subgroup 1 is a Braille document created on plain paper without an overlaid pattern. There are related research topics such as Braille character extraction for Braille document images recognition [1–6], Braille document images recognition [7–12], Parameter estimation of a Braille document images [13, 14], and Recovering the Braille embossed point of an old Braille document [15].

When considering the document image processing diagram for Braille documents as shown in Fig. 1 and comparing it with Fig. 2. In Fig. 2, the area to the right of the red dash line is a new research topic that has not been researched previously. This research project deals with Braille embossed dot segmentation for Braille document images produced on reusable paper. It has an overlaid pattern, as shown in Fig. 3. Therefore, this research is classified in Subgroup 2 of Group 2.

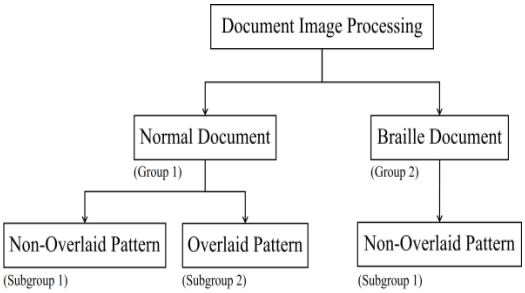


Fig. 1. The Diagram of Document Image Processing.

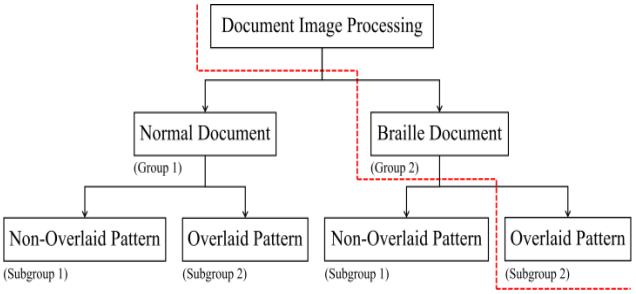


Fig. 2. The Diagram of the Document Image Processing with the Area to the Right of the Red Dash Line is a Research Topic that has not been previously Researched.

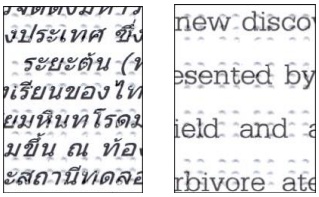


Fig. 3. Example of Braille Document Images Produced on Reusable Paper.

Subgroup 1 of Group 2 is related research: C. M. Ng et al. [1] used the Chain Code and grouped the Braille embossed dots into cells. A. Al-Saleh et al. [2], Braille document images modeled using Beta distribution for thresholding and used grid for Braille embossed dots segmentation. A. S. Al-Salman et al. [3] uses image enhancement and image rotation. Then a grid method is used to extract the Braille Cells. A.-S. Amany et al. [4] uses thresholding based on Beta distribution then creates a grid for dot detection. M. Y. Babadi et al. [5] perform skew correction and create grids for Braille cells segmentation. A. AlSalman et al. [6], this research use between-class variance with Gamma distribution to separate Braille embossed dots from the background. J. Mennens et al. [7], Braille documents are imported with a scanner. The mask and grid are used to extract the Braille embossed dots. L. Wong et al. [8], this research use techniques Half-Character Recognition then generate a grid to extract Braille embossed dots. L. Jie et al. [9, 10], this research uses Support Vector Machine (SVM), slides window techniques, and Haar wavelet to extract Braille embossed dots on Braille document images obtained from a scanner. B.-M. Hsu [11], this research used RCSA: Ratio Character Segmentation Algorithm for Braille embossed dots extraction. A. AlSalman et al. [12], the research use the Deep Convolution Neural Network (DCNN) for Braille document recognition. M. Yousefi et al. [13, 14], this research finds the parameters of Braille documents, skewness, scaling, line spacing to obtain Braille dots. H. Kawabe et al. [15] uses deep learning to classify Braille dots in long-preserved or ancient Braille documents. It can be seen that these studies focus on Braille documents created on plain paper only. There has been no research that has created Braille on reusable paper.

III. Proposed Method

This research proposed the method of Braille embossed dots segmentation for Braille document images. Braille documents were produced on reusable paper and plain paper. Flatbed scanners were used to scan the documents. The method comprises six steps as shown in Fig. 4. The first step was to perform a scanner calibration process using a specific calibration plate and calculating the edges' threshold values and the black areas. The second step was to perform a layer separation process by using the edges' threshold values and the black areas from the previous step. The third step was to perform the edge removal process by eliminating the edges. The fourth step was to perform the data recovery process by calculating the eroded mask's data to increase the Braille dots' details. The next step was to perform a noise removal process by applying an explosion algorithm to diffuse the black pixels and then to image enhancement by spatial filtering. The final process was to perform a Braille localization process by calculating the Braille dots' positions from the black pixels' positions. The details are as follows.

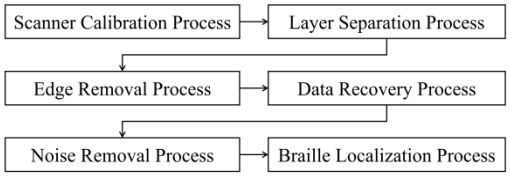


Fig. 4. The Diagram of Proposed Method.

A. Scanner Calibration Process

1) Plates for scanner calibration: The images were designed for scanner calibration, a black circle, and a black square then saved to the image files. Then printed on plain paper by using a laser printer and an inkjet printer is called the plates. Then, they were imported via a flatbed scanner.

2) The plates were converted to grayscale images: The plates were in the RGB color model. They were converted to the HSV color model. The color values of the V plane were applied to all three planes of the RGB color model. The grayscale images are shown in Fig. 5(a) and 5(b), comprising the black area in P-01 and the background in P-02.

3) Plate images are eroded with a mask: They were converted to binary images and then eroded with a small mask and large mask, respectively. As a result, it can be seen that the plate images were eroded by the small mask have a black area in P-03 of Fig. 5(c) and 5(d), are larger than the plate images eroded by the large mask, which have a black area in P-04 of Fig. 5(e) and 5(f).

4) Threshold values of the edges calculation: The plate images eroded by a small mask were compared with the plate image eroded by a large mask. The location of pixels with different color values in P-05 of Fig. 5(g) and 5(h), the color value of the grayscale image in Fig. 5(a) and 5(b) in that the position were stored in the array $Edge$, the threshold value for the edges of images was calculated by using the equation:

$$T1_{Edge} = round((Edge_{mean} + Edge_{\min})/2) \quad (1)$$

$$T2_{Edge} = round(Edge_{mean} + sedge) \quad (2)$$

$$sedge = abs((Edge_{\max} - Edge_{mean})/2) \quad (3)$$

where $T1_{Edge}$, $T2_{Edge}$ are the first and the second threshold values for the edges of images, respectively. $Edge_{mean}$, $Edge_{\min}$ and $Edge_{\max}$ are mean, min and max values of the grayscale color, respectively.

5) Threshold values of the black areas calculation: The plate images eroded by a small mask were compared with the image eroded by a large mask. The location of pixels with the same color values in P-06 of Fig. 5(g) and 5(h), the color value of the grayscale image of Fig. 5(a) and 5(b) in that the position were stored in the array $Black$, the threshold value of black areas was calculated by using the equation:

$$T1_{Black} = round(Black_{mean} + sblack) \quad (4)$$

$$sblack = abs((Black_{\max} - Black_{mean})/2) \quad (5)$$

where $T1_{Black}$ is the threshold values for the black area of images, $Black_{mean}$ and $Black_{\max}$ are mean and max values of the grayscale color, respectively.

B. Layer Separation Process

1) Braille document images importing: A Braille document was imported by using the calibrated scanner from the previous step. They were color images in the RGB model converted to grayscale, the same as in step 2) of the previous process. Fig. 6(a) contains the background of the Braille document in P-01, the Braille embossed dot in P-02, and the typical character that was called a pattern shows in P-03.

2) Edges image calculation: The color values of the grayscale image in the pixel positions were compared with the threshold value of the edges derived from Equations (1) to (3). If the color values of the grayscale image in the pixel positions are between $T1_{Edge}$ and $T2_{Edge}$ values, black color values were recorded in the white image for the edges image at the same pixel positions. The result was called the image of the edges, as shown in Fig. 6(b). It included the edges of the pattern shows in P-04, and the Braille embossed dots in P-05. There may be noise shown in P-06.

3) Black areas image calculation: The color values of the grayscale image in the pixel positions were compared with the threshold value of the black areas derived from Equation (4) and (5). If the color values of the grayscale image in the pixel positions were less than $T1_{Black}$ values, black color values were recorded in the white image for the black area image at the same pixel positions. The result was called the image of the black areas, as shown in Fig. 6(c). It included the black areas of the pattern shows in P-07 and the holes of punctures in P-08.

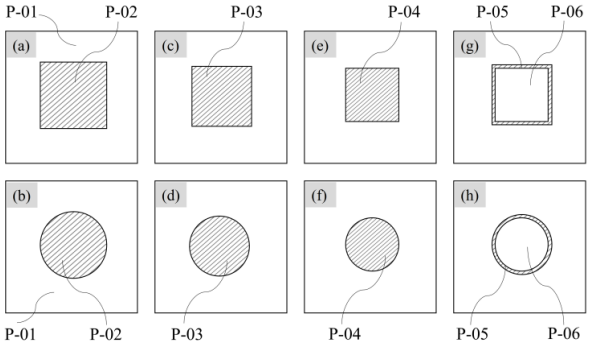


Fig. 5. Illustrations of the Scanner Calibration Process.

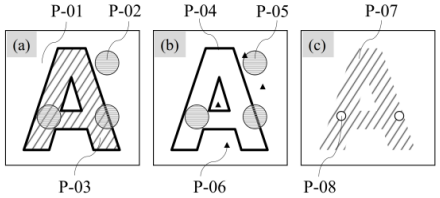


Fig. 6. Illustrations of the Layer Separation Process.

C. Edge Removal Process

1) Image of the black areas dilation: The image results from the previous process as shown in Fig. 7(a). The black area's image was inverted color, the result is shown in Fig. 7(b). The black areas of the pattern were dilated with a mask. Then, the black areas that were larger and the holes of the puncture are filled, therefore the image of the black areas dilated, as shown in Fig. 7(c) and P-03 was larger than P-01 in Fig. 7(b).

2) Edges removed: The OR logic operation of the dilated black areas image shown in Fig. 7(c) and the image of the edges obtained in the previous process, Fig. 7(d), which can remove the edges of the pattern. In Fig. 7(e), the result is the image of the edges removed that still has Braille embossed dots in P-04 and P-06 and some noise in P-05, as shown in Fig. 7(e).

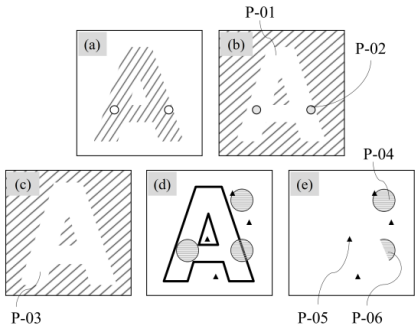


Fig. 7. Illustrations of the Edge Removal Process.

D. Data Recovery Process

1) Image of the black area's erosion: The inverted image of the black areas dilated from the previous process, Fig. 8(a), were eroded with a mask. In Fig. 8(b), the white areas are smaller than before. Fig. 8(c) showed the inverted color, called the image of the eroded black areas, which shown in P-01.

2) The braille dots recovery: The OR logic operation of the image of the black areas eroded, shown in Fig. 8(c), and the image of the edges, shown in Fig. 8(d), which can recover the details of the Braille embossed dot. And then, it was combined with the image of the edges, as shown in Fig. 8(e). The result was the image of removed edges and the details, as shown in Fig. 8(f), P-02 is the Braille embossed dot and P-03 is some noise.

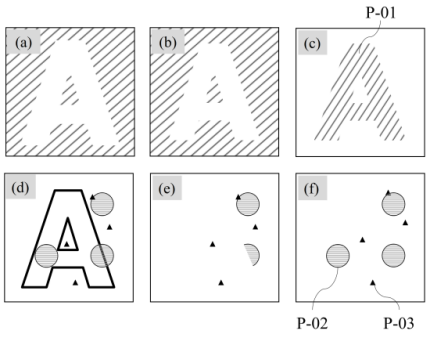


Fig. 8. Illustrations of the Data Recovery Process.

E. Noise Removal Process

1) Diffusion of the edges: The image of removed edges and details, Fig. 9(a), is exposed by a mask. The mask moves one pixel at a time to determine the color value. If the mask's center position was black value, it would be moved to a new position inside the mask. The black value with the new position was saved in the white image data. The result was the image of diffused edges as shown in Fig. 9(b), P-01 to P-03 are diffused dot and some noise.

2) Image enhancement: The images of diffused edges were applied by averaging filter, which created a clearer Braille embossed dot image. The result was an enhanced image as shown in Fig. 9(c).

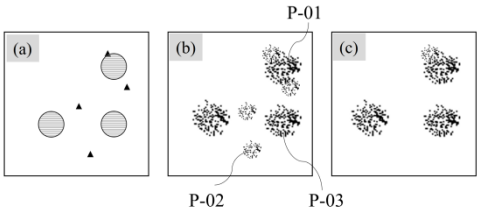


Fig. 9. Illustrations of the Noise Removal Process.

F. Braille Localization Process

1) Braille embossed dot dilation: The enhanced image, Fig. 10(a), was dilated by a mask. The result was a group of black values that were Braille embossed dots for more clarity as shown in Fig. 10(b).

2) The centroid of braille embossed dot: The group of black values in the previous step was calculated by connected component labeling. The result was the centroid of Braille embossed dot, which was a Braille embossed dots location as shown in P-01 of Fig. 10(c).

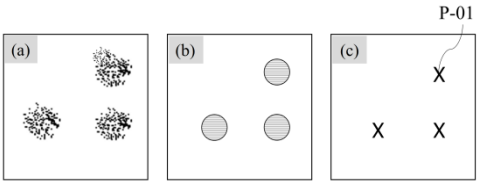


Fig. 10. Illustrations of the Braille Localization Process.

IV. Experiments and Results

A. Dataset

This research produced a dataset of Braille embossed dots named KU-Braille-Dot. Braille documents were produced on reusable paper by using slate and stylus. Consider Table I, this research created reusable paper by using an 80 GSM A4 paper to print text by using a LaserJet printer and Inkjet printer. It was then used to create Braille documents by using the slate and stylus. These Braille documents were scanned by using four scanners, each with different lighting environments, as shown in Fig. 11. The Braille document images had a resolution of 300 DPI. It is cropped to size 40×40 pixels under six events; each event contains 100 image data obtained from 50 images from the Inkjet printer and 50 images from LaserJet printer. The details of the six events were: (1) Event 1: Braille

embossed dots were in the middle of any text, 100 images. (2) Event 2: Braille embossed dots overlap any text, but they were visible, 100 images. (3) Event 3: Braille embossed dots were on a plain background, 100 images. (4) Event 4: Braille embossed dots overlap any text, but they were not visible, 100 images. (5) Event 5: Plain background with no Braille embossed dot and no text, 100 images. (6) Event 6: Plain background with text and no Braille embossed dot, 100 images.

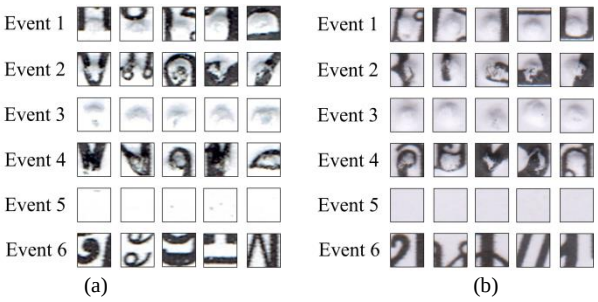
B. Experimental Design

The proposed method was tested by creating a single-cell Braille with various patterns from the KU-Braille-Dots dataset. A single-cell Braille contains six dot positions, each of which can occur in six events, so 6×6×6×6×6×6 is equal to 46,656 patterns in total. A dot position of Braille has six events. The events were arranged in order, and each event was a randomized image from 100 images, called a data group in the A form, as shown in Fig. 12. A single-cell Braille contained six dots, and therefore required six groups of data in the A form, which were sorted into a data group, called the B form, as shown in Fig. 13. Each row was a group of data in the A form relative to the dot positions of a single-cell Braille. A group of data in the B form which could create a single-cell Braille with 46,656 patterns.

This research created 150,000 groups of data in the A form. Each group had a unique event image or could contain no more than two duplicate event images. Those data were then randomly grouped into 262,144 groups in the B form, and each group had a unique arrangement of event images. That is, 262,144 cells Braille, which had 12,230,590,464 patterns.

TABLE I. DETAILS OF THE KU-BRAILLE-DOT DATASET

Scanner	Event	Number of Images	Scanner	Event	Number of Images
Light Condition 1	Event 1	100	Light Condition 2	Event 1	100
	Event 2	100		Event 2	100
	Event 3	100		Event 3	100
	Event 4	100		Event 4	100
	Event 5	100		Event 5	100
	Event 6	100		Event 6	100
Light Condition 3	Event 1	100	Light Condition 4	Event 1	100
	Event 2	100		Event 2	100
	Event 3	100		Event 3	100
	Event 4	100		Event 4	100
	Event 5	100		Event 5	100
	Event 6	100		Event 6	100



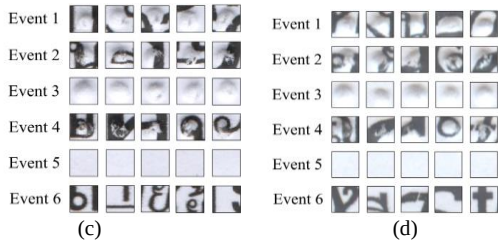


Fig. 11. Example of the KU-Braille-Dots: (a) Light Condition 1 (b) Light Condition 2 (c) Light Condition 3 and (d) Light Condition 4.

Event 01	Event 02	Event 03	Event 04	Event 05	Event 06
35	32	4	96	21	57

Fig. 12. An Example of the Data Group in a Form and the Index of Event Images.

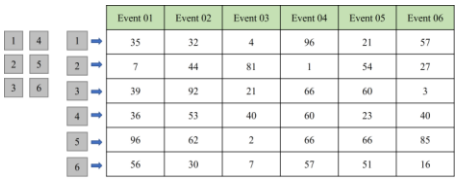


Fig. 13. A Group of Data in b Form.

TABLE II. THE NUMBER OF BRAILLE CELL INCREASES BY 1 TIMES

Group No.	Braille Cell	Group No.	Braille Cell
1	1	11	1,024
2	2	12	2,048
3	4	13	4,096
4	8	14	8,192
5	16	15	16,384
6	32	16	32,768
7	64	17	65,536
8	128	18	131,072
9	256	19	262,144
10	512		

C. Performance Measurement

This research was to test the performance measurement of the proposed method using the number of the cell Braille increased by 1 times, starting from a single-cell Braille up to 262,144 cells Braille, as shown in Table II. The performance measurement of the proposed method, a single-cell Braille was used to describe, as shown in Fig. 14. The details were: (1) Consider the area No. 1, the gray area is the Braille embossed dots, and the white area is the background. (2) The ground truth of a single-cell Braille is shown in the area No. 3, the green area is the Braille embossed dots, and the blue area is the background. (3) In the area No. 2, the results are Braille embossed dots obtained from the proposed method. Positions 2, 3 and 5 are Braille embossed dots and the positions 1, 4 and 6 are non-Braille embossed dots or background. (4) Consider the results obtained from the proposed method and the ground truth. In area No. 4, positions 4 and 5 answered incorrectly. (5) This was used to evaluate the performance measurement of the proposed method by calculating Precision, Recall, Accuracy, and F-Measure, respectively.

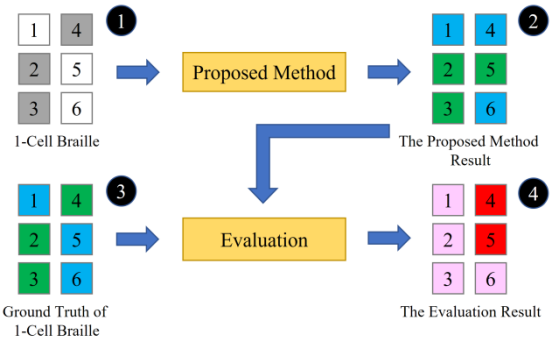


Fig. 14. The Diagram of the Performance Measurement.

D. Results and Discussion

The presented method was tested with the data sets described in the previous section. This research aimed to extract the Braille embossed dots from Braille documents created on reusable paper. An example of the calculation according to the proposed method using a single Braille embossed dot is shown in Fig. 15. These documents were scanned by using four flatbed scanners, videlicet, four lighting conditions. This research plotted graphs of the light condition as shown in Fig. 16 to 19, each of which has four lines representing Precision, Recall, Accuracy, and F-Measure, respectively, and has a value between 0.00 and 1.00. The y-axis of the graph is the numerical measure, which is a value between 0.00 and 1.00, but these graphs start plotting at a value of 0.6 for clarity. The x-axis of the graph is group No. of the Braille datasets used for testing.

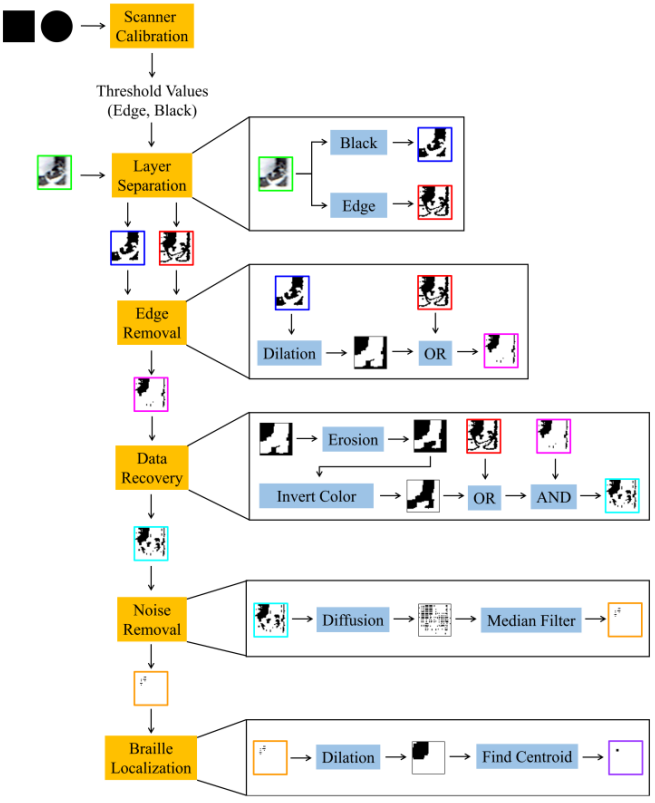


Fig. 15. The Schematic Diagram describes the Processing according to the Proposed Method.

Consider Fig. 16 to 19; in the Precision lines, the four lighting conditions have a Precision value of 1.00. In the Recall lines, light conditions 1, 2, and 4 are approximately 0.80, and light condition 3 is approximately 0.70. The Accuracy lines in all light conditions are over 0.80. In the F-measure lines calculated from Precision and Recall, each lighting condition is over 0.80. When calculating the average performances in all lighting conditions, the values are Precision 1.0000, Recall 0.7817, Accuracy 0.8545, and F-Measure 0.8756.

By considering the Accuracy and the F-measure values greater than 0.80, it is known that there is approximately one position dot error in a single-cell Braille with six-position dots. This research is to extract Braille embossed dots on reusable paper. It is different from other research [1-15], which is only interested in Braille documents on plain paper.

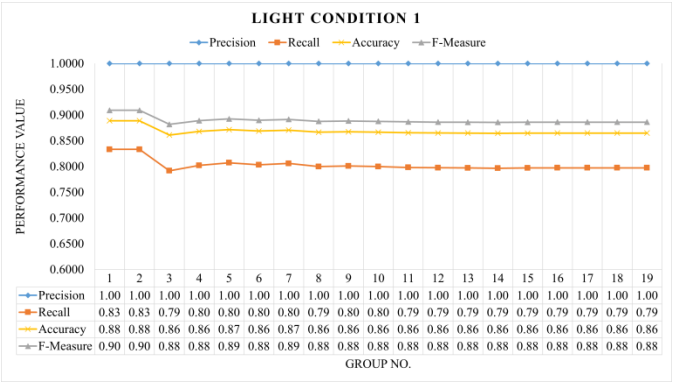


Fig. 16. The Performance Result for the Dataset in Light Condition 1.

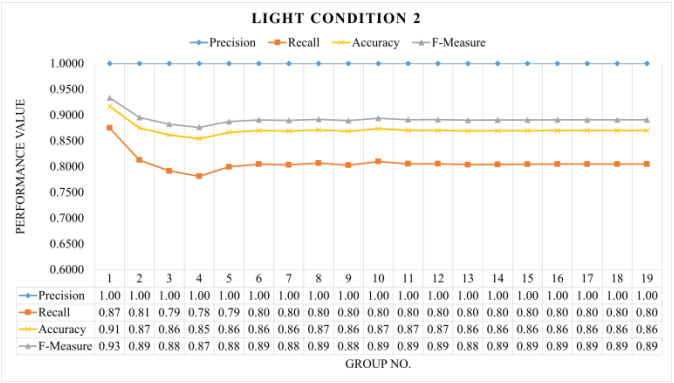


Fig. 17. The Performance Result for the Dataset in Light Condition 2.

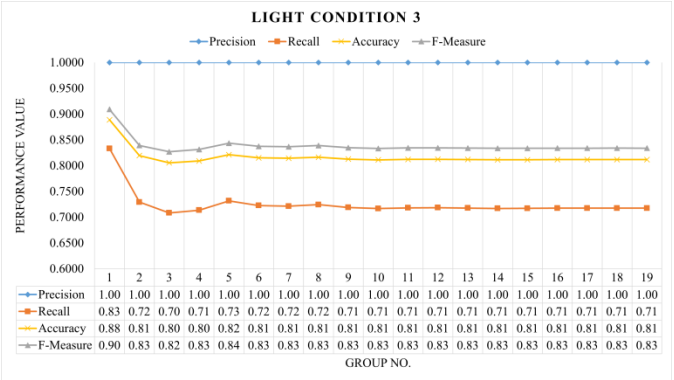
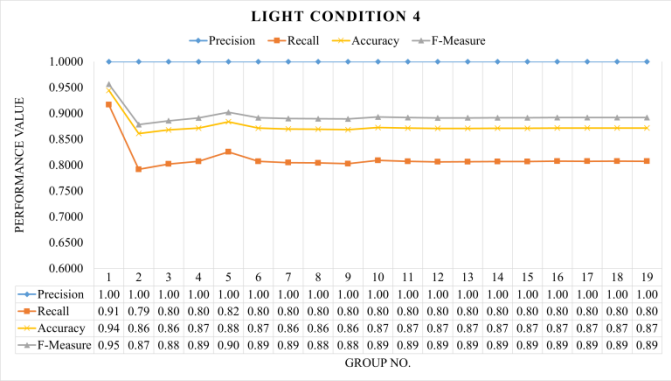


Fig. 18. The Performance Result for the Dataset in Light Condition 3.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	0.91	0.79	0.80	0.80	0.82	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Accuracy	0.94	0.86	0.86	0.87	0.88	0.87	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
F-Measure	0.95	0.87	0.88	0.89	0.90	0.89	0.89	0.88	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89

Fig. 19. The Performance Result for the Dataset in Light Condition 4.

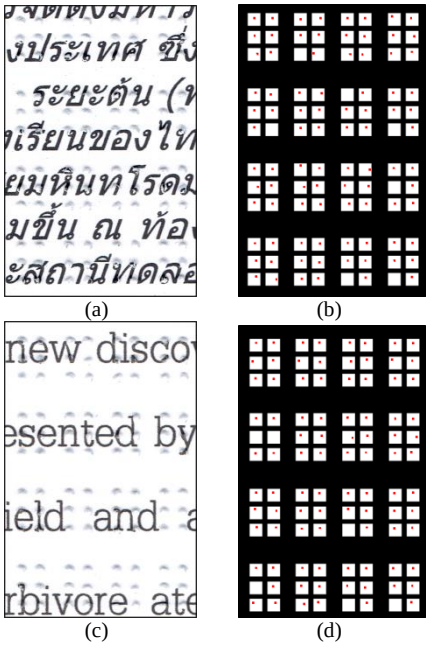


Fig. 20. Preliminary Experimental Results for the Proposed Method on KU-Braille-Partial: (a) and (c) were Parts of a Braille Document Image. (b) and (d) Resulted from the Proposed Method.

V. Conclusion

This paper presented the Braille embossed dots segmentation method for Braille document images that were tested as detailed. The results were good, with an average Accuracy and F-Measure of over 0.85 for all lighting conditions. This research confirmed that the proposed method has a good efficiency for Braille embossed dot segmentation for Braille documents produced on reusable paper, positively affecting the Braille recognition method.

The research included diverse flatbed scanner devices and lighting conditions. In this field of research, no research has been found on Braille documents produced on reusable paper. The methods presented here make the best use of paper resources and increase access to education for the blind. This research has opened a research path that is beneficial to the visually impaired or blind people to have more opportunities to study and learn.

VI. Future Work

This research is a starting point to help create a digital document from documents created by the blind by using a slate and a stylus on reusable paper. It also facilitates communication between ordinary people and the blind or visually impaired and promotes the desire to be treated as ordinary. In the future, this method will be tested on the KU-Braille-Partial dataset to achieve higher accuracy. Further developed methods are applied to actual Braille documents created on reusable paper. The KU-Braille-Partial was a dataset of parts of a Braille document image produced on reusable paper using slate and stylus. This research created reusable paper using an 80 GSM A4 paper to print text using a LaserJet printer and Inkjet printer. They were scanned by using four scanners with a resolution of 300 DPI. It was cropped to size 585x915 pixels—the sample images as shown in Fig. 20(a) and 20(c). The ground truth images were created—the sample images as shown in Fig. 20(b) and 20(d). The white rectangular areas were the Braille embossed dots, and the black areas were the non-Braille embossed dots. The red dot in the white square means that the proposed method was correct, but in other cases, it was wrong. The results obtained from this preliminary experiment showed that the proposed method was practical and that the research scale could be scaled up.

Acknowledgment

The authors wish to thank the many people for their contribution to this project; Ms. Chavee Tiendee, Ms. Chatchada Nanto, Ms. Pranom Sirimangkalo, Ms. Benjamas Toumbumrung, and Ms. Kwanruethai Naksongkaew, for their help in creating the Braille document dataset. Special thanks should be given to Kasetsart University for the Ph.D. scholarship.

References

[1] C. M. Ng, V. Ng, and Y. Lau, "Regular feature extraction for recognition of Braille," in Proceedings Third International Conference on Computational Intelligence and Multimedia Applications. ICCIMA'99 (Cat. No.PR00300), 23-26 Sept. 1999 1999, pp. 302-306, doi: 10.1109/ICCIMA.1999.798547.

[2] A. Al-Saleh, A. El-Zaart, and A. Alsalman, "Dot Detection of Optical Braille Images for Braille Cells Recognition," presented at the Proceedings of the 11th international conference on Computers Helping People with Special Needs, linz, Austria, 2008. [Online]. Available: https://doi.org/10.1007/978-3-540-70540-6_122.

[3] A. S. Al-Salman, A. El-Zaart, Y. Al-Suhaibani, K. Al-Hokail, and A. O. Al-Qabbany, "An Efficient Braille Cells Recognition," in 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 23-25 Sept. 2010 2010, pp. 1-4, doi: 10.1109/WICOM.2010.5601020.

[4] A.-S. Amany, E.-Z. Ali, and A.-S. Abdul Malik, "Dot Detection of Braille Images Using A Mixture of Beta Distributions," Journal of Computer Science, vol. 7, no. 11, 09/06 2011, doi: 10.3844/jcssp.2011.1749.1759.

[5] M. Y. Babadi and S. Jafari, "Novel grid-based optical Braille conversion: from scanning to wording," International Journal of Electronics, vol. 98, no. 12, pp. 1659-1671, 2011/12/01 2011, doi: 10.1080/00207217.2011.609975.

[6] A. AlSalman, A. El-Zaart, S. Al-Salman, and A. Gumaei, "A novel approach for Braille images segmentation," in 2012 International Conference on Multimedia Computing and Systems, 10-12 May 2012 2012, pp. 190-195, doi: 10.1109/ICMCS.2012.6320146.

[7] J. Mennens, L. V. Tichelen, G. Francois, and J. J. Engelen, "Optical recognition of Braille writing using standard equipment," IEEE

Transactions on Rehabilitation Engineering, vol. 2, no. 4, pp. 207-212, 1994, doi: 10.1109/86.340878.

[8] L. Wong, W. Abdulla, and S. Hussmann, "A software algorithm prototype for optical recognition of embossed Braille," in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 26-26 Aug. 2004 2004, vol. 2, pp. 586-589 Vol.2, doi: 10.1109/ICPR.2004.1334316.

[9] L. Jie and Y. Xiaoguang, "Optical Braille character recognition with Support-Vector Machine classifier," in 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 22-24 Oct. 2010 2010, vol. 12, pp. V12-219-V12-222, doi: 10.1109/ICCASM.2010.5622245.

[10] L. Jie, Y. Xiaoguang, and Z. Dayong, "Optical Braille recognition with Haar wavelet features and Support-Vector Machine," in 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, 24-26 Aug. 2010 2010, vol. 5, pp. 64-67, doi: 10.1109/CMCE.2010.5610062.

[11] B.-M. Hsu, "Braille Recognition for Reducing Asymmetric Communication between the Blind and Non-Blind," Symmetry, vol. 12, no. 7, p. 1069, 2020. [Online]. Available: https://www.mdpi.com/2073-8994/12/7/1069.

[12] A. AlSalman, A. Gumaei, A. AlSalman, and S. Al-Hadhrami, "A Deep Learning-Based Recognition Approach for the Conversion of Multilingual Braille Images," Computers, Materials \& Continua, vol. 67, no. 3, 2021, doi: 10.32604/cmc.2021.015614.

[13] M. Y. Babadi, B. Nasihatkon, Z. Azimifar, and P. Fieguth, "Probabilistic estimation of Braille document parameters," in 2009 16th IEEE International Conference on Image Processing (ICIP), 7-10 Nov. 2009 2009, pp. 2001-2004, doi: 10.1109/ICIP.2009.5413816.

[14] M. Yousefi, M. Famouri, B. Nasihatkon, Z. Azimifar, and P. Fieguth, "A robust probabilistic Braille recognition system," International Journal on Document Analysis and Recognition (IJDAR), vol. 15, no. 3, pp. 253-266, 2012/09/01 2012, doi: 10.1007/s10032-011-0171-7.

[15] H. Kawabe, Y. Shimomura, H. Nambo, and S. Seto, "Application of Deep Learning to Classification of Braille Dot for Restoration of Old Braille Books," in Proceedings of the Twelfth International Conference on Management Science and Engineering Management, Cham, J. Xu, F. L. Cooke, M. Gen, and S. E. Ahmed, Eds., 2019// 2019: Springer International Publishing, pp. 913-926.

[16] A. Delaye and C.-L. Liu, "Contextual text/non-text stroke classification in online handwritten notes with conditional random fields," Pattern Recognition, vol. 47, no. 3, pp. 959-968, 2014/03/01/ 2014, doi: https://doi.org/10.1016/j.patcog.2013.04.017.

[17] S. S. Bukhari, Z. Iftikhar, and A. Dengel, "Chemical Structure Recognition (CSR) System: Automatic Analysis of 2D Chemical Structures in Document Images," in 2019 International Conference on Document Analysis and Recognition (ICDAR), 20-25 Sept. 2019 2019, pp. 1262-1267, doi: 10.1109/ICDAR.2019.00-41.

[18] B. H. Phong, T. M. Hoang, and T. L. Le, "A Hybrid Method for Mathematical Expression Detection in Scientific Document Images," IEEE Access, vol. 8, pp. 83663-83684, 2020, doi: 10.1109/ACCESS.2020.2992067.

[19] S. Elaiwat and M. Abu-Zanona, "Arabic Word Recognition System for Historical Documents using Multiscale Representation Method," International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, 2020 2020, doi: 10.14569/IJACSA.2020.01104107.

[20] D. P. Van Hoai, H.-T. Duong, and V. T. Hoang, "Text recognition for Vietnamese identity card based on deep features network," International Journal on Document Analysis and Recognition (IJDAR), vol. 24, no. 1, pp. 123-131, 2021/06/01 2021, doi: 10.1007/s10032-021-00363-7.

[21] A. M.Hesham, S. Abdou, A. Badr, M. Rashwan, and H. M.Al-Barhamtoshy, "A Zone Classification Approach for Arabic Documents using Hybrid Features," International Journal of Advanced Computer Science and Applications, vol. 7, no. 7, 2016 2016, doi: 10.14569/IJACSA.2016.070722.

[22] J. Calvo-Zaragoza, K. Zhang, Z. Saleh, G. Vigliensoni, and I. Fujinaga, "Music Document Layout Analysis through Machine Learning and Human Feedback," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 9-15 Nov. 2017 2017, vol. 02, pp. 23-24, doi: 10.1109/ICDAR.2017.259.

[23] S. T. R. Rizvi, A. Dengel, and S. Ahmed, "A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction," IEEE Access, vol. 8, pp. 217231-217245, 2020, doi: 10.1109/ACCESS.2020.3042455.

[24] R. Arief, A. B. Mutiara, T. M. Kusuma, and Hustinawaty, "Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment," International Journal of Advanced Computer Science and Applications, vol. 9, no. 11, 2018 2018, doi: 10.14569/IJACSA.2018.091117.

[25] Y. Liu, C. Si, K. Jin, T. Shen, and M. Hu, "FCENet: An Instance Segmentation Model for Extracting Figures and Captions From Material Documents," IEEE Access, vol. 9, pp. 551-564, 2021, doi: 10.1109/ACCESS.2020.3046496.

[26] A. Zhu, C. Zhang, Z. Li, and S. Xiong, "Coarse-to-fine document localization in natural scene image with regional attention and recursive corner refinement," International Journal on Document Analysis and Recognition (IJDAR), vol. 22, no. 3, pp. 351-360, 2019/09/01 2019, doi: 10.1007/s10032-019-00341-0.

[27] D.-D. Nguyen, "TableSegNet: a fully convolutional network for table detection and segmentation in document images," International Journal on Document Analysis and Recognition (IJDAR), 2021/11/22 2021, doi: 10.1007/s10032-021-00390-4.

[28] P. D. Ingle and P. Kaur, "Adaptive thresholding to robust image binarization for degraded document images," in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 5-6 Oct. 2017 2017, pp. 189-193, doi: 10.1109/ICISIM.2017.8122172.

[29] F. Drira and F. LeBourgeois, "Mean-Shift segmentation and PDE-based nonlinear diffusion: toward a common variational framework for foreground/background document image segmentation," International Journal on Document Analysis and Recognition (IJDAR), vol. 20, no. 3, pp. 201-222, 2017/09/01 2017, doi: 10.1007/s10032-017-0285-7.

[30] M. Hanif, A. Tonazzini, P. Savino, and E. Salerno, "Non-Local Sparse Image Inpainting for Document Bleed-Through Removal," Journal of Imaging, vol. 4, no. 5, p. 68, 2018. [Online]. Available: https://www.mdpi.com/2313-433X/4/5/68.

[31] D. Rodin, A. Zharkov, and I. Zagaynov, "Faster Glare Detection on Document Images," in Document Analysis Systems, Cham, X. Bai, D. Karatzas, and D. Lopresti, Eds., 2020// 2020: Springer International Publishing, pp. 161-167.

[32] J. R. Wang and Y. Y. Chuang, "Shadow Removal of Text Document Images by Estimating Local and Global Background Colors," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-8 May 2020 2020, pp. 1534-1538, doi: 10.1109/ICASSP40776.2020.9053378.

[33] C.-M. Tsai, "Intelligent region-based thresholding for color document images with highlighted regions," Pattern Recogn., vol. 45, no. 4, pp. 1341–1362, 2012, doi: 10.1016/j.patcog.2011.09.024.

Implementation of Password Hashing on Embedded Systems with Cryptographic Acceleration Unit

Holman Montiel A, Fredy Martínez S, Edwar Jacinto G
Facultad Tecnológica
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

***Abstract*—In this modern world where the proliferation of electronic devices associated with the Internet of Things (IoT) grows day by day, security is an imperative issue. The criticality of the information linked to the various electronic devices connected to the Internet forces developers to establish protection mechanisms against possible cyber-attacks. When using computer equipment or servers, security mechanisms can be applied without having problems with the number of resources associated with this activity; the opposite is the case when implementing such mechanisms on embedded systems. The objective of this document is to implement password hashing on a FRDM-K82F development board with ARM® Cortex™-M4 processor. It describes the basic criteria necessary to aim at moderate levels of security in specific purpose applications; that can be developed taking advantage of the hardware cryptographic acceleration units that these embedded systems have. Performance analysis of the implemented hash function is also presented, considering the variation in the number of iterations performed by the development board. The validation of the correct functioning of the hashing scheme using the SHA-256 algorithm is carried out by comparing the results obtained in real-time versus an application developed in Python software using the PyCryptodome library.**

***Keywords*—Cryptography; password hashing; embedded systems; cryptographic acceleration hardware; SHA-256**

I. Introduction

One of the current priorities with the boom and great demand for devices associated with the Internet of Things (IoT) [1], in the face of multiple interconnectivity environments is security [2],[3]. While it is true that day by day the development of specific solutions or electronic control units associated with communication processes; establishes a great demand by this society interconnected to the web [4]; as developers, it must be considered that there are a wide variety of tools both software and hardware [5]; which allow to improve and optimize security against the handling of critical information [6],[7]. The possible vulnerability and impact on the integrity of the information make security a key point in any electronic development; thus, establishing a primordial factor in the selection criteria of the possible users of these technological solutions [8].

A fundamental characteristic that must be considered when implementing IoT on embedded systems is that most of these devices have limitations associated with computational power and speed [9], [10]. Although this could be a limitation when developing electronic control units with cryptographic functions [11]; many applications are being developed that make use of hardware coprocessors that facilitate the implementation of various cryptographic algorithms and hash functions [12], [13]. This allows the use of these embedded systems to be applied not only in encryption and decryption tasks [14], but also in hashing and authentication in complex environments with limited computing resources [15],[16]. The current literature shows us a great variety of security implementations on embedded systems in which performance analysis is performed for both symmetric and asymmetric algorithms [17], [18]; this shows us that cryptographic hardware continues in a constant process of evolution due to the great demand for efficient [19], reliable, portable [20], and secure IoT technological products [21].

In this complex scenario of IoT interconnectivity, some organizations have fallen prey to cyber-attacks in which they have been exposed thanks to vulnerabilities exploited by poor password protection practices. The exploitation of this vulnerability has managed to expose many user accounts and credentials; significantly affecting the reliability of the use of web applications and embedded solutions used in home automation and industrial areas [6],[12],[14]. One way to counteract this phenomenon and at the same time guarantee a high level of security related to user accounts is through the implementation of password hashing schemes (PHS) [13].

Considering the above, this work aims to show the implementation and validation of a password hashing on the FRDM-K82F embedded system. The document intends to develop in a simple and practical way a first approach to the use of Arm Mbed TLS libraries; thus, providing a basic and functional solution for those who make their first approach to the use of processors with cryptographic acceleration units. A performance analysis associated with the variation in the number of iterations used for the generated summary function is also shown.

This contribution is presented in the following sections, which are organized as follows: Section II describes the basic theoretical concepts associated with the use of embedded systems with cryptographic acceleration unit. Section III describes the implementation and development of the proposal. Section IV presents the results obtained, and finally, the conclusions and future work are presented in Section V.

II. Methodology

The constant evolution of digital devices in terms of their cryptographic modules has allowed a lot of opportunities;

associated with the linking of security parameters in the development of specific purpose applications. Using the FRDM-K82F development platform, a step-by-step implementation of password hashing is carried out, making use of the Cryptographic Acceleration Unit (CAU). The validation and verification of the SHA-256 algorithm and the respective results are performed by means of the Python PyCryptodome library; this to verify the correct operation of the proposed solution on the embedded system used.

A. Cryptographic Acceleration Unit

A wide variety of embedded systems can be found in the market that has this type of module incorporated in their architecture, see Fig. 1. In general, terms, the Cryptographic Acceleration Unit (CAU) is a ColdFire® coprocessor that is accessed by the CPU using specialized hardware operations [21], [22]. The purpose of this unit is to increase the performance of software-based hashing and encryption functions, thus guaranteeing acceleration and high performance when using algorithms such as DES, 3DES, AES, MD5, SHA, among others.

Also available are some Kinetis® MCU processors that have the memory-mapped cryptographic acceleration unit (mmCAU), a coprocessor that is connected to the processor's private peripheral bus (PPB), as shown in Fig. 2. These units are focused on improving the performance of software-based security encryption/decryption operations.

B. Password Hashing

One way to increase security in any communication and information transfer process associated with applications that link user accounts with their respective access passwords; is to move from storing passwords in plain text to using a hash function on the respective password, as shown in Fig. 3. A hash function makes it possible to encrypt a password by taking advantage of the fact that this type of algorithm; takes any size of data and converts it into a fixed length of information [13]. To be more precise, the aim is to make it impossible to recover the password from the generated hash.

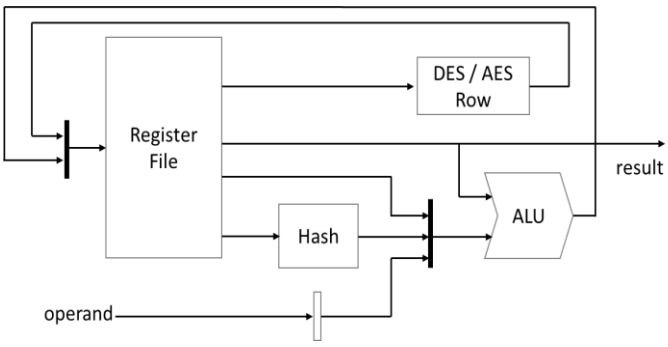


Fig. 1. Block Diagram of the CAU Module [22].

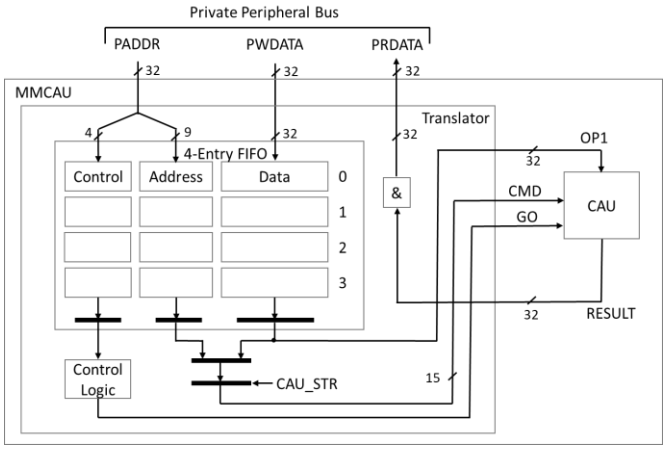


Fig. 2. mmCAU Block Diagram [23].

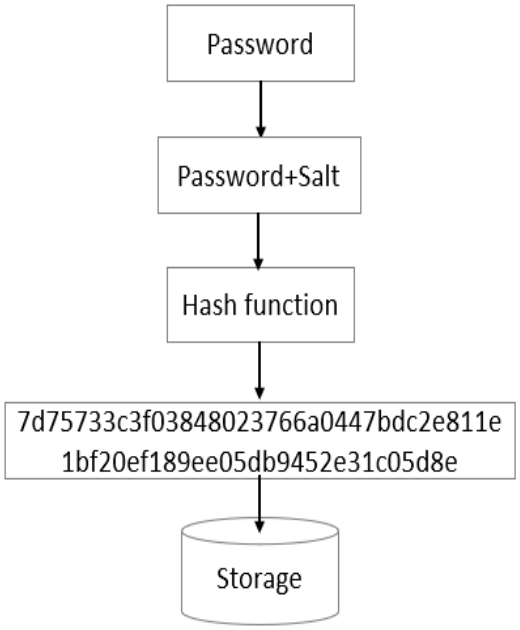


Fig. 3. Suggested Password Storage.

As a good security practice, it is not recommended to store passwords in unencrypted text; since any attacker could access them and obtain all the information directly. By storing the hash of these passwords and taking advantage of the fact that these functions are not reversible, the security vulnerability of the respective system is significantly reduced. The validation process of this technique is simple; initially, the input data is taken, the same hash function is executed, and then it is analyzed if this result matches the information stored in the password store, see Fig. 4.

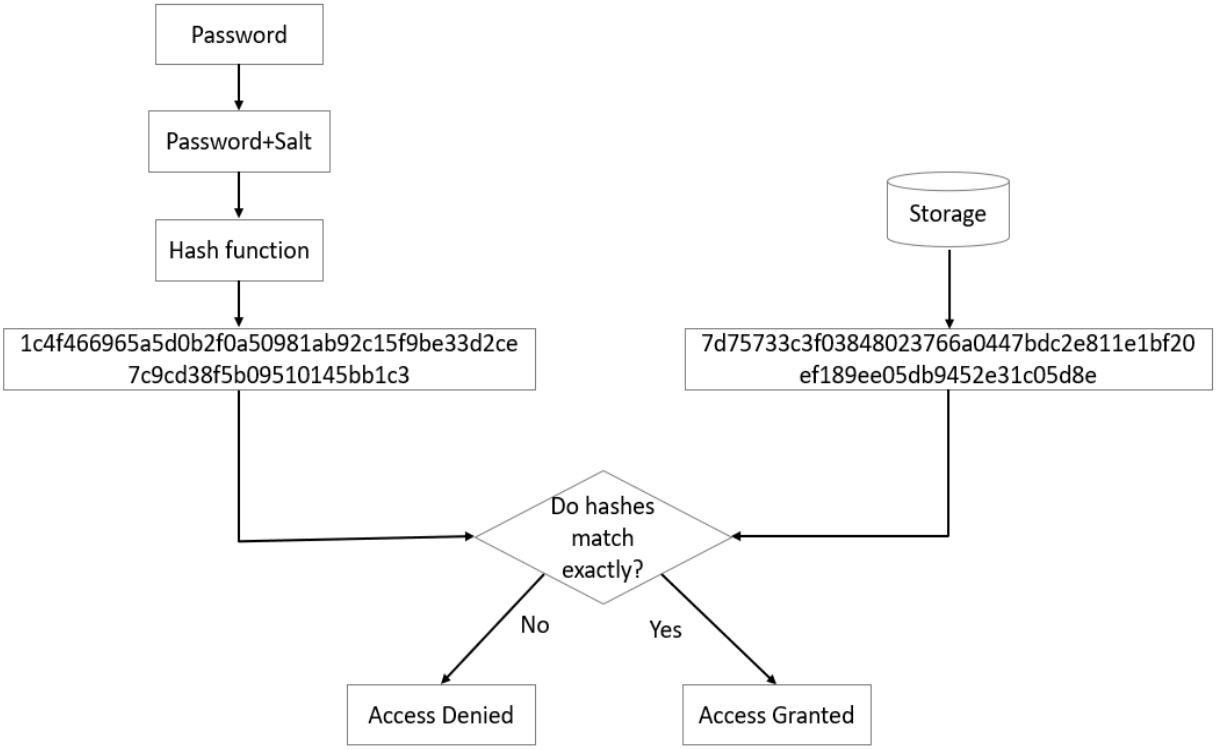


Fig. 4. Password Validation.

III. Implementation

Initially, use is made of the PyCryptodome libraries, which is an autonomous package of low-level cryptographic primitives supported by Python. This software tool allows an external validation of the operation of the cryptographic process to be implemented. It can be said that first the operational validation of the process will be performed by means of software simulation; to later compare the results obtained with the implementation on hardware with the FRDM-K82F platform, see Fig. 5.

scratch.py

```
import time as t
from Crypto.Hash import SHA256



t1 = t.time()
sal = '5m8k@q1$'
datos = 'passwordHMA'
hash_object = SHA256.new(data=bytes(datos+sal))
#print(hash_object.hexdigest())

for _ in range(1,100000):
    hash_object = SHA256.new(data= bytes(hash_object.digest()))

t2 = t.time() - t1
print('P :', hash_object.hexdigest() )
print('time = ', t2)

```

Fig. 5. Source Code used in Python for Validation of Results.

As the fundamental objective of the application is to make use of the cryptographic acceleration unit (CAU) of the FRDM-K82F card; which has an ARM® Cortex®-M4 core running at up to 150 MHz, with KB256 of Flash and 256 KB of RAM. The source code is developed using the Mbed OS compiler, which provides a comprehensive SSL/TLS solution called Arm MbedTLS [24]. This library simplifies the integration of cryptographic solutions because it is compact and generic; it should be noted that it can only be used in ColdFire and Kinetis devices with CAU or mmCAU hardware coprocessors. The following encryption/decryption algorithms and hash functions can be used with this library: AES128, AES192, AES256, DES, 3DES, MD5, SHA1, and SHA256.

This implementation was carried out using the Mbed-Studio compiler; this has a large repository of examples associated with the security schemes [25]. It must be considered that at the moment of creating the source code and loading the libraries; the compiler must initially evaluate if the processor has the cryptographic acceleration unit required for the use of the respective libraries; based on this concept, some components of the source code would be as follows:

#include "mbed.h"

#include "mbedtls/sha256.h" /* SHA-256 only */

#include "mbedtls/md.h" /* generic interface */

#include < cstdio>

#if DEBUG_LEVEL > 0

#include "mbedtls/debug.h"

```
#endif

#include "mbedtls/platform.h"

#include < string.h>
```

The call to the Hash function is quite simple, the definition of the respective input and output variables must be considered; for this case it must be considered that the output associated to a SHA256 function is of 32 bytes, with respect to the input it must be remembered that it can be of any length. The parameters and input variables would be:

```
static const char dato_input[] = "passwordHMA5m8k@q1$";

static const unsigned char *input_buffer = (const unsigned
char *) dato_input;

static const size_t dato_len = strlen(dato_input);
```

With respect to the execution of the SHA function, four functional requirements must be considered. These are: Data buffer, buffer length, Output buffer and a parameter defining whether to use the full SHA-256 or the SHA-224 variant. In this case, this value must be 0 (to use SHA-256).

```
unsigned char output1[32]; /* SHA-256 outputs 32 bytes */

unsigned char output2[32];

t.start();

mbedtls_sha256(input_buffer, dato_len, output1, 0);

for (int i=1; i<=99999; i++){

mbedtls_sha256(output1, 32, output1, 0);

  }

  t.stop();
```

IV. Results

To validate the effectiveness of the password hashing implementation on the selected hardware, we proceeded to compare the result obtained with the implementation done entirely on the PC using pyCryptodome, (see characteristics in Table I). The execution times were measured both in the PC implementation and in the embedded system, performing a variation between the number of iterations associated with the selected hash function.

As general concepts, a test password of 11 bytes in length was used, encrypted with a SHA256 algorithm (password + salt). The salt used was eight bytes long. For the information associated with the salt, a test constant was used and there was no code segment associated with its generation by the embedded system. The validation and verification were performed by comparing the output of the simulation performed in Python versus the result given by the serial port of the development board. Comparisons of up to 100,000 iterations were performed, demonstrating the full functionality of the implementation, see Fig. 6 and Fig. 7.

TABLE I. Characteristics of the Devices Used

Characteristics of the processors used		
Platform:	**Embedded System**	**PC**
Reference:	FRDM-K82F	ROG GL553VD
Processor:	Kinetis MK82FN256VLL15 (ARM® Cortex™-M4)	Intel® Core™ i7-7700HQ
Clock frequency:	150 MHz	2.8 GHz
RAM:	256 KB	12 GB

Fig. 6. Comparison of Software vs. Hardware Results.

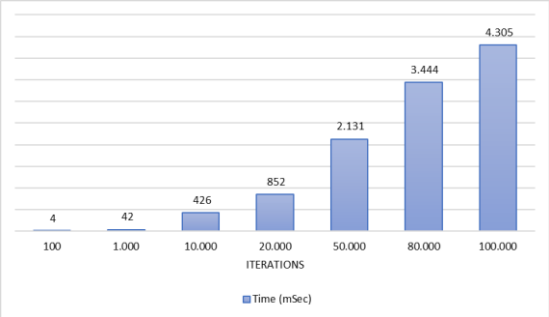


Fig. 7. Execution Time on the FRDM-K82F Board.

V. Conclusion and Future Work

This document presents simply and easily a password hashing scheme that can be implemented on any embedded system with cryptographic acceleration hardware; it provides the key concepts so that people who are starting in the subject of embedded cryptography can begin to incorporate these security measures for their respective developments. It shows that it is possible to make use of a technological solution that offers a moderate level of security using electronic devices with limited computational resources. Although in most IoT applications, communication between devices is short term and there is not a high rate of information transfer; the protection of session passwords becomes a fundamental objective in terms of security. It is hoped that this type of examples will encourage developers to incorporate new methodologies for protecting information on the design of IoT devices.

As future work, we intend to develop specific application hardware to enable digital signatures by implementing hash functions and lightweight encryption algorithms.

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 the implementation on hardware.

References

[1] S. Surendran, A. Nassef and B. D. Beheshti, "A survey of cryptographic algorithms for IoT devices," *2018 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1-8, 2018.

[2] B. Vinayaga Sundaram, Ramnath M., Prasanth M. and Varsha Sundaram J., "Encryption and hash based security in Internet of Things," *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, pp. 1-6, 2015.

[3] M. El-Haii, M. Chamoun, A. Fadlallah and A. Serhrouchni, "Analysis of Cryptographic Algorithms on IoT Hardware platforms," *2018 2nd Cyber Security in Networking Conference (CSNet)*, pp. 1-5, 2018.

[4] P. Flood and M. Schukat, "Peer to peer authentication for small embedded systems: A zero-knowledge-based approach to security for the Internet of Things," *The 10th International Conference on Digital Technologies 2014*, pp. 68-72, 2014.

[5] C. Profentzas, M. Günes, Y. Nikolakopoulos, O. Landsiedel and M. Almgren, "Performance of Secure Boot in Embedded Systems," *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 198-204, 2019.

[6] R.V. Rashmi and A. Karthikeyan, "Secure boot of Embedded Applications - A Review," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 291-298, 2018.

[7] A. Mundra, A., & H. Guan, "Secure Boot on Embedded Sitara Processors". *Texas Instrument*, 2018.

[8] N. B. Silva, D. F. Pigatto, P. S. Martins, & K.C.Branco, "Case studies of performance evaluation of cryptographic algorithms for an embedded system and a general purpose computer", *Journal of Network and Computer Applications*, Vol. 60, pp. 130-143, 2016.

[9] N. J. G. Saho, & E. C. Ezin, "Survey on Asymmetric Cryptographic Algorithms in Embedded Systems", *IJISRT*, vol *5*, no. 12, pp. 544-554, 2020.

[10] A. J. Acosta, T. Addabbo, & E. Tena-Sánchez, "Embedded electronic circuits for cryptography, hardware security and true random number generation: an overview", *International Journal of Circuit Theory and Applications*, vol. *45, no.* 2, pp.145-169, 2017.

[11] Z. Musliyana, T. Y. Arif, & R. Munadi, "Security enhancement of advanced encryption standard (AES) using time-based dynamic key generation", *ARPN Journal of Engineering and Applied Sciences*, vol *10, no.* 18, pp. 8347-8350, 2015.

[12] W. Wang et al. "XMSS and Embedded Systems". In: Paterson K., Stebila D. (eds) Selected Areas in Cryptography – SAC 2019. SAC 2019. Lecture Notes in Computer Science, vol 11959. Springer, Cham, pp. 1-33, 2020.

[13] G. Hatzivasilis, I. Papaefstathiou, C. Manifavas, I. Askoxylakis, "Lightweight Password Hashing Scheme for Embedded Systems", In: Akram R., Jajodia S. (eds) Information Security Theory and Practice. WISTP 2015. Lecture Notes in Computer Science, vol 9311. Springer, Cham, 2015.

[14] A. Flores-Vergara, E. Inzunza-González, E. E. García-Guerrero, O. R. López-Bonilla, E. Rodríguez-Orozco, J. M. Hernández-Ontiveros, E. Tlelo-Cuautle, "Implementing a chaotic cryptosystem by performing parallel computing on embedded systems with multiprocessors", *Entropy*, vol. *21, no.* 3, pp. 1-28, 2019.

[15] M. Mozaffari-Kermani, K. Tian, R. Azarderakhsh and S. Bayat-Sarmadi, "Fault-Resilient Lightweight Cryptographic Block Ciphers for Secure Embedded Systems," in *IEEE Embedded Systems Letters*, vol. 6, no. 4, pp. 89-92, Dec. 2014.

[16] P. Branco, L. Fiolhais, M. Goulão, P. Martins, P. Mateus, and L. Sousa, "ROTed: Random Oblivious Transfer for embedded devices", *TCHES*, vol. 2021, no. 4, pp. 215–238, Aug. 2021.

[17] L. Baldanzi, L. Crocetti, F. Falaschi, M. Bertolucci, J. Belli, L. Fanucci, and S. Saponara, "Cryptographically Secure Pseudo-Random Number Generator IP-Core Based on SHA2 Algorithm" *Sensors*, *vol.* 20, no. 7, 1869, pp. 1-13, 2020.

[18] S. Falas, C. Konstantinou and M. K. Michael, "A Hardware-based Framework for Secure Firmware Updates on Embedded Systems," *2019 IFIP/IEEE 27th International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 198-203, 2019.

[19] Z. He, W. Chen, X. Xu, L. Harn, & M. Wan, "Reliable and efficient PUF-based cryptographic key generator using bit self-tests", *Electronics Letters*, vol. *56, no.*16, pp. 803-806, 2020.

[20] Z. Gu, G. Han, H. Zeng and Q. Zhao, "Security-Aware Mapping and Scheduling with Hardware Co-Processors for FlexRay-Based Distributed Embedded Systems," in IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 10, pp. 3044-3057, 1 Oct. 2016.

[21] K. Q. Ye, M. Green, N. Sanguansin, L. Beringer, A. Petcher, and A. W. Appel, "Verified Correctness and Security of mbedTLS HMAC-DRBG". *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, 2017.

[22] L. Casado, "Soluciones de seguridad de Freescale parte III: aceleradores criptográficos en la familia de procesadores Coldfire", *Revista española de electrónica*, (649), pp. 76-80, 2008.

[23] Freescale, "Kinetis MCUs Securing the Internet of Tomorrow", Rev2, pp. 1-12, 2015.

[24] ARM mbed,"mbed TLS v2.16.1 source code documentation", 2021.

[25] ARM mbed, "Tutorial and official examples", Repository Mbed OS 6, 2021.

Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers

Ertel Merouane[1]
Informatics and Applications Laboratory (IA), Faculty of Sciences
Moulay Ismail University
Meknes, Morocco

Amali Said[2]
Informatics and Applications Laboratory (IA), FSJES
Moulay Ismail University
Meknes, Morocco

El Faddouli Nour-eddine[3]
RIME Team, MASI Laboratory
E3S Research Center, EMI
Mohammed V University
Rabat, Morocco

***Abstract*—The volume and amount of data in cancerology is continuously increasing, yet the vast majority of this data is not being used to uncover useful and hidden insights. As a result, one of the key goals of physicians for therapeutic decision-making during multidisciplinary consultation meetings is to combine prediction tools based on data and best practices (MCM). The current study looked into using CRISP-DM machine learning algorithms to predict metastatic recurrence in patients with early-stage (non-metastatic) breast cancer so that treatment-appropriate medicine may be given to lower the likelihood of metastatic relapse. From 2014 to 2021, data from patients with localized breast cancer were collected at the Regional Oncology Center in Meknes, Morocco. There were 449 records in the dataset, 13 predictor variables and one outcome variable. To create predictive models, we used machine learning techniques such as Support Vector Machine (SVM), Nave Bayes (NB), K-Nearest Neighbors (KNN) and Logistic Regression (LR). The main objective of this article is to compare the performance of these four algorithms on our data in terms of sensitivity, specificity and precision. According to our results, the accuracies of SVM, kNN, LR and NB are 0.906, 0.861, 0.806 and 0.517 respectively. With the fewest errors and maximum accuracy, the SVM classification model predicts metastatic breast cancer relapse. The unbiased prediction accuracy of each model is assessed using a 10-fold cross-validation method.**

***Keywords*—Machine learning; classification; personalized medicine; CRISP-DM; metastasis; breast cancer**

I. Introduction

Breast cancer is a significant public health concern. According to data released by the World Cancer Observatory in 2018, 52,783 new cancer cases are reported in Morocco each year, with women accounting for 36.9% of these cases [1] , The key events linked to poor survival in breast cancer patients are disease progression and metastasis. Adjuvant chemotherapy (treatment given after surgery) combined with hormone therapy has been demonstrated in some trials to minimize the risk of recurrence and mortality from breast cancer [2], [3]. Due to the development of metastases and uncontrolled growth, various cases of female patients do not respond to therapeutic compounds in breast cancer in the same way [4].

Over the past two decades, personalized medicine has been defined in several ways. More broadly as a predictive, personalized, preventive and participatory health model ("P4 medicine") [5], and which also applies technologies to personalize and deliver care [6]. The use of personalized medicine or precision medicine in oncology aims to adapt treatments according to the characteristics of patients and their diseases by integrating all the biological and genetic, environmental, phenotypic and psychosocial knowledge found there clean [7]. Personalized medicine's ultimate goal is to provide the appropriate treatment to the appropriate person at the appropriate time [8].

The statistical method of machine learning techniques has shown to be a godsend for diagnostic, classification, prediction, and prognosis purposes in personalized medicine in cancer, given the amount of clinical data about each patient [9]–[13]. Various researchers are applying machine learning ideas to enhance cancer prediction and prognosis, this is done using a training data set whose variable assignments are already predetermined or known. Recently, researchers have focused more on decision trees, KNNs, SVMs and neural networks to predict cancer patient survival with high accuracy [14]–[16] . Web-based prediction models have been developed from cancer registry data to help determine the need for adjuvant therapy [17], [18]. PREDICT uses multivariate statistical analysis to calculate personalized survival probability based on the integration of clinical factors [19], [20]. However, the use of these models in clinical practice relies heavily on proof of the reliability of predictions and demonstration of acquired knowledge, moreover, the majority of them focus on overall survival rather than the risk of relapse. Given the paucity of predictive machine learning models that allow clinicians to identify patients at risk for metastatic relapse earlier by using a combination of various clinic-pathological characteristics, in particular Ki67 with tumor size, lymph node invasion and adjuvant therapy, we have seen fit to continue the current effort to resolve this problem.

In this study, our objective is to propose a supervised learning model, for predicting metastatic recurrence in individuals with early-stage breast cancer on an individual basis, which will guide the therapeutic decision in the multidisciplinary consultation meeting (MCM). Our model is fed by data including clinical, pathological, biological, therapeutic and prognostic characteristics. These data are collected from the files of patients with early-stage breast cancer, collected after the different stages of treatment

(diagnosis, relapse/progression, follow-up), offering a holistic view of previous successes and recommendations for good practices.

In the second part of this article, we will present the predictor variables introduced into the model, which predict the risk of metastatic relapse in patients before the start of adjuvant treatments (chemotherapy - Hormone therapy - Radiotherapy - Trastuzumab). The model proposal obtained according to the CRISP-DM process will be presented in the third section and in the last section we will analyze the results.

II. Related Work

In medical practice, the efficiency of breast cancer treatment is essentially determined on the ability to cancer prognosis, and cancer recurrence [21]. In recent years, with the use of machine learning technology in personalized medicine [6], modern oncology seeks to tailor treatments to expected results, through personalized predictive care models, based on patient characteristics patients and their pathologies by integrating all the biological and genetic, environmental, phenotypic and psychosocial knowledge that are specific to it. Tseng and Yi-Ju (2019) [22] propose an approach based on machine learning such as Random Forest (RF), Support Vector Machine (SVM), logistic regression (LR) and Naive Bayes (NB), to predict early breast cancer metastases using serum biomarkers and clinicopathologic data to reduce the risk of death. Tapak and Leili (2019) [23] proposed a model based on learning algorithms such as Naive Bayes (NB), Random Forest (RF), AdaBoost, Support Vector Machine (SVM), Least-squareSVM (LSSVM), Adabag, Logistic Regression (LR) and Linear Discriminant Analysis (LDA), for the prediction of breast cancer survival and metastasis.

In our research, we investigated four machine learning methods for predicting metastases in breast cancer patients: Support vector machine, Naive Bayes, K-Nearest Neighbour, and Logistic Regression. These algorithms are integrated into our proposed model according to the standard CRISP-DM process. The description of this model is the subject of the following section.

III. Materials and Methods

The phase of creating a predictive machine learning model is preceded by a preprocessing phase. In order to feed the model with clean data, the data may contain values that need to be transformed or eliminated, which can be useful for modeling. In our study, The CRISP-DM approach was employed (Cross Industry Standard Process for Data Mining) [24] which is considered to be an essential pillar for the success of a Machine Learning (ML) project. This method can help us find information and patterns hidden in a dataset with many features [25]. The CRISP method has six phases (see Fig. 1) that we will detail in the sections.

A. Data Understanding

1) Data source: Our predictive study included patients with localized breast cancer on all histological types of cancer collected at the regional oncology center of Meknes in Morocco, during the period 2014 to 2021, who had undergone surgery associated with adjuvant treatment during the years 2014 - 2016 (Chemotherapy and / or Hormonetherapy and / or Radiotherapy and / or Trastuzumab) with a follow-up of at least 48 months.

Our system's dataset contains 511 records and 14 variables. These variables provide demographic, clinical and therapeutic information about the patient, including the target variable (metastatic relapse). The data were collected from the computerized system which brings together the archives of patient files, which were then validated by experts (treating physicians).

2) Dataset features: The decision for adjuvant systemic treatment of breast cancer is based on clinical factors, such as (age) and histological (axillary lymph nodes, size, grade, vascular emboli), performance indicators, previous treatment methods, but also organic [26] Co-morbidities and of course the wishes of patients will continue to play an important role. To determine the adjuvant treatment of non-metastatic breast cancer [27]. Biologically, the expression of HR (Hormone Receptors) and the overexpression of HER2 (human epidermal growth factor receptor 2) are the main prognostic biomarkers and key predictors of the therapeutic effect [28], [29]. However, other biological parameters seem to have emerged recently, a study showed that the Ki67 index of cell proliferation in univariate and multivariate analyses of grade cancers, was the strongest predictor of overall and metastasis-free survival [30]. It is an important biomarker in the management of breast cancer, it can be used to guide clinical decisions regarding adjuvant chemotherapy [31].

The variables collected from the computer system of the regional oncology center of Meknes-Morocco, were the prognostic factors currently validated in breast cancer : The age of the patient, the size of the tumor, the pathological state of the lymph nodes lymphatic, grade, stage, histological type of tumor, estrogen receptor (ER) status, progesterone receptors (PR) grouped into hormone receptors (HR), HER2 overexpression, Ki67 status (cell proliferation), L ' surgical approach and types of adjuvant therapy.

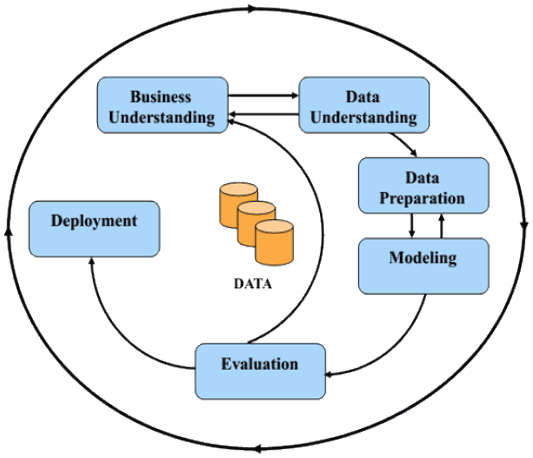


Fig. 1. Phases of the Current CRISP-DM Process Model for Data Mining

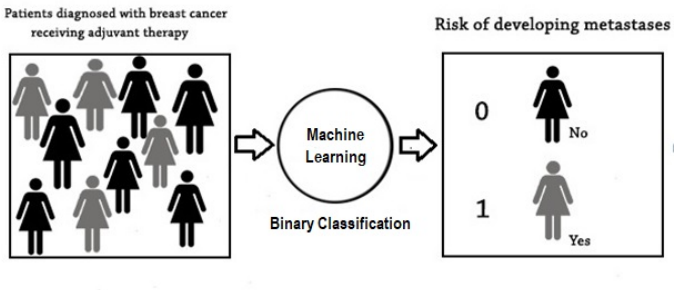


Fig. 2. Our Binary Classification Model for the Prediction of Metastatic Recurrence.

Our predictive model is based on these indicators to classify new patients with non-metastatic breast cancer into two classes: patients at low (0) or high (1) risk of metastatic relapse at 4 years (see Fig. 2).

B. Data Preparation

Data preparation is made up of several stages: Data cleaning, Data Transformation.

1) Data cleaning: The data collected from the information system of the Meknes Regional Oncology Center in Morocco is organized in the form of a database. This database has undergone a cleaning process to eliminate and reduce noise:

- Attribute noise is caused by input errors, missing variable values and redundant data.
- Class noise which is due to errors introduced when assigning instances to classes.

After removing the rows with substantial missing values, we checked for missing or null data points in the database using Python's pandas library (see Fig. 3).

The number of records kept is 449 records, each showing a different case of breast cancer with its own combination of treatments. Each of these cases is represented by 13 independent predictors / variables, plus 1 dependent / categorical variable that reflects metastatic relapse in breast cancer patients (No / Yes).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 449 entries, 0 to 448
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age_diagnosis       449 non-null    int64
 1   Tumor_size          449 non-null    int64
 2   Lymph_Nodes         449 non-null    int64
 3   Tumor_stage         449 non-null    int64
 4   Cancer_Grade        449 non-null    int64
 5   HER2                449 non-null    object
 6   HR                  449 non-null    object
 7   Ki67                449 non-null    int64
 8   Surgery_Type        449 non-null    object
 9   Chemotherapy        449 non-null    object
 10  Trastuzumab         449 non-null    object
 11  Radiotherapy        449 non-null    object
 12  Hormonotherapy      449 non-null    object
 13  Metastatic_Relapse  449 non-null    object
dtypes: int64(6), object(8)
memory usage: 49.2+ KB
```

Fig. 3. Attribute Information of the Dataset.

2) Data transformation: The quality of the data and the amount of useful information are key factors that determine the learning ability of a machine learning algorithm. Therefore, it is absolutely essential to make sure that we encode categorical variables correctly, before using the data in a machine learning algorithm [32]. In this study we have 14 distinct attributes: 3 attributes represent numeric characteristics, 10 attributes represent object type variables, and the last attribute represents an object output variable, this means that our data contains object / categorical type variables, they must be coded by numbers before we can fit and evaluate our model.

For this, we used the technique (OneHotEncoder) from the Scikit-Learn library in the Pandas module of Python, to create a hot-encoding of integer-encoded values, which transforms the input categorical variables into numbers. This method increases the overall number of input characteristics, so this type of encoding creates a binary variable for each unique value of the nominal characteristic. The binary variable specifies (0) or (1) whether or not the category appears in observation (see Table I).

C. Modeling

The data preprocessing step is followed by a modeling process, which involves training the machine learning algorithms to predict the classes from the features. In this study, the presented entries are normalized so that all variables are on the same scale and distribution, in order to compare the performance of the models and evaluate them in the same way. We used the method (model_selection.KFold) of the SciKit-Learn library in Python, to train the model to create the cross-validation folds by 10. Indeed, this method is used to evaluate predictive models which divide the set original into a training sample that represents the training DataSet, and another set reserved for testing and evaluating the model. The result is a trained model that can be used for inference; making predictions on new data points (see Fig. 4).

TABLE I. CHARACTERISTICS OF PATIENTS WITH BREAST CANCER AND POSSIBLE VALUES

Attribute	***Type Attribute***	***Possible Value***
Age_diagnosis	*Numerical*	*20 - 80 (Years)*
Tumor_size	*Numerical*	*10 mm - 70 mm*
Lymph_Nodes	*Categorical*	*0 - 3*
Tumor_stage	*Categorical*	*0 - 3*
Cancer_Grade	*Categorical*	*1 - 3*
HER2	*Categorical*	*0 (Negative) - 1 (Positive)*
HR	*Categorical*	*0 (Negative) - 1 (Positive)*
Ki67	*Numerical*	*8 % - 60 %*
Surgery_Type	*Categorical*	*0 (Tumorectomy) - 1 (Mastectomy)*
Chemotherapy	*Categorical*	*0 (No) - 1 (Yes)*
Trastuzumab	*Categorical*	*0 (No) - 1 (Yes)*
Radiotherapy	*Categorical*	*0 (No) - 1 (Yes)*
Hormonotherapy	*Categorical*	*0 (No) - 1 (Yes)*
Metastatic_Relapse	*Categorical*	*0 (No) - 1 (Yes)*

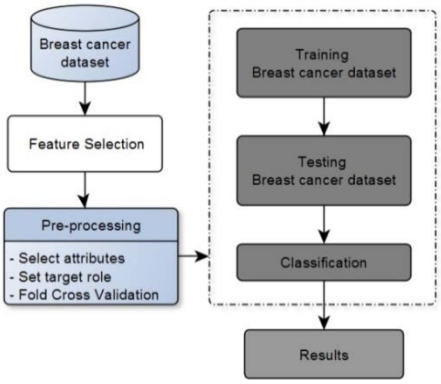


Fig. 4. Supervised Learning Workflow.

1) Classification methods: In the present study, four machine learning methods were used and compared to predict metastasis in breast cancer patients: Support Vector Machine, Naive Bayes, K-Nearest Neighbors, and Logistic Regression.

a) Support Vector Machine (SVM): SVM are a type of supervised learning algorithms for classification, regression analysis, and outlier identification that examine data [33]. It's a discriminating model described by a hyperplane; in our case, the hyperplane categorizes new instances into one of two classes: 0 or 1.

b) Naïve Bayes (NB): The influence of a variable value on a specific class is assumed to be independent of the values of other variables by NB classifiers. This is referred to as conditional class independence. When the training dataset is small, it is utilized to identify crucial classification parameters [33]. The NB classifier, which combines the Bayes probability model with a decision rule, is one of the most extensively used binary classification algorithms.

c) K-Nearest Neighbors (KNN): The non-assumption of the variable's distribution is one of the method's advantages. When comparing the two preceding techniques, this is a crucial consideration. To maximize classification and cope with the bias-variance trade-off, this approach must determine the optimal value of k, the number of neighbors. Optimal choices of k keep the bias-variance balance in check and, ideally, reduce both [34].

d) Logistic Regression (LR): LR is a classification algorithm generally used in binary classification problems [35], as is the case here with negative, 0 and positive response values, 1. It uses the maximum likelihood estimate for assess the probability of belonging to a class.

2) Performance measures: It is necessary to calculate the model's accuracy in order to test its capacity to anticipate occurrences in the proper class. The following procedures were employed.

a) Confusion Matrix: This is a statistic for evaluating a classification model's performance. It's also known as an error matrix since it may be used to figure out where the model is off in its predictions. The confusion matrix analyzes the number of accurate and wrong predictions after the prediction. On the basis of these factors, classifier comparisons are made (see Fig. 5).

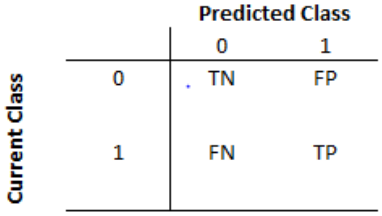


Fig. 5. Confusion Matrix.

We may designate one class as positive and one as negative per row and true or false per column in binary classification, giving us:

- TP: correct relapse expected.
- TN: correction of the expected non-relapse.
- PF: incorrectly predicted relapse.
- FN: incorrect non-relapse prediction.

b) Classification report: A classification report is used to assess the classification model's quality.

The proportion of right guesses in the overall number of correct forecasts is known as accuracy. It is calculated by (1), where TP and TN indicate the number of properly categorized positive and negative cases, respectively, and FN and FP represent the number of incorrectly classified negative and positive examples.

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

The ratio of true positives to all positives is called precision (2). This would be the measurement of individuals accurately identified as having a risk of metastatic recurrence among all patients truly at risk for our issue statement.

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

Equation (3) defines the true negative rate (specificity). Among all negative data points, the false positive rate is the fraction of negative data points that are correctly classified as negative.

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

The real positive rate, calculated by equation (4), is the recall (sensitivity). Out of all positive data points, this rate represents the proportion of positive data points that are accurately classified as positive.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

The Roc curve and AUC: When the decision threshold is changed, a Receiver Operating Characteristic (ROC) curve displays the rate of true positives (sensitivity) versus the rate of false positives (1 - specificity) [36]. The area under the curve (AUC) is a measure of the likelihood that the model would rate a positive random example higher than a negative random example. Its values range from 0 to 1. The AUC of a model with 100% incorrect predictions is 0. Its AUC is 1 if all of its predictions are right.

The comparison of the performance of learning algorithms, discussed in the next section, is based on these indicators (Accuracy; Precision; specificity; recall; AUC).

IV. Results and Discussion

We utilized Jupyter Notebook, Python modules (pandas, matplotlib, bumpy), and the scikit-learn framework to process ML algorithms for our analysis. To predict metastasis in breast cancer patients, the following approaches were tested: (SVM, NB, k-NN, and LR).

First, we performed training for 70% of the dataset (314 random records), applying the cross-validation method checking all the metrics mentioned previously. Then we ran a test of the remaining 30% of the data set. Table II illustrates the prediction results by successfully classified and wrongly categorized examples for the methods (SVM, NB, kNN, and LR).

Then we compared the difference between the precision results found in the test and the total, this comparison is based on the indicators Accuracy, Precision, sensitivity, specificity, Roc curve and AUC, to measure the performance of these algorithms based on the Confusion Matrix entries. The findings are shown in Table III.

The best classification performance is obtained with SVM, as shown in Table II, which correctly predicts 123 instances out of 135 (94 instances 0 which are in fact 0 and 29 instances 1 which are in fact 1), and 12 badly predicted instances (03 instances of class 0 predicted as 1 and 09 instances of class 1 predicted as 0). We also notice that NB has the lowest value of correctly classified instances and the highest value of misclassified instances (36 badly predicted instances) compared to the other classifiers (12 incorrect instances for kNN and LR).

In Table III, the results of the performance measurements of the four classification algorithms clearly show that the SVM and kNN achieved the highest precision (91.1%). kNN has reached the highest sensitivity (Recall), which is 81.6%. And NB the worst specificity (51.7%). We can also notice that SVM surpasses the other classifiers in terms of Precision (90.6%), Specificity (96.9%), AUC (92.9%). This is why, with a score of (91.1%) and a smaller error, the SVM outperforms the other classification approaches utilized in our study.

TABLE II. Confusion Matrix of Classification Techniques BLE

Classifiers	***Predicted***		***Test Size = 0.30***	
	0	***1***		
SVM	*94*	*3*	***0***	***Current***
	9	*29*	***1***	
kNN	*92*	*5*	***0***	
	7	*31*	***1***	
LR	*90*	*7*	***0***	
	9	*29*	***1***	
NB	*69*	*28*	***0***	
	8	*30*	***1***	

TABLE III. Classifiers Performance

Classifiers	***Accuracy (%)***	***Precision***	***Specificity***	***Recall***	***AUC***
SVM	*91,1*	*0,906*	*0,969*	*0,763*	*0,929*
kNN	*91,1*	*0,861*	*0,948*	*0,816*	*0,882*
LR	*88,1*	*0,806*	*0,928*	*0,763*	*0,914*
NB	*73,3*	*0,517*	*0,711*	*0,789*	*0,750*

The ROC curve, on the other hand, offers for a better grasp of a machine learning algorithm's capability. Fig. 6 shows the ROC curves displayed for the fitted test models in our investigation.

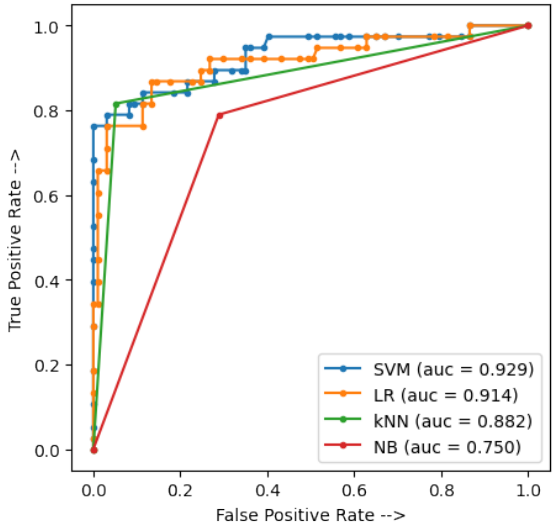


Fig. 6. Roc Curve (AUC) Models.

SVM is the best classifier, as seen in Fig. 6, since the curve is squished towards the top left edge, then travels towards the upper right corner (90.6 % sensitive and 96.9% specific), followed by the other algorithms: LR, kNN, and NB.

The results obtained for various performance indicators show the effectiveness of the SVM model in predicting metastatic relapse in patients with early-stage breast cancer with the highest precision value of 91.1% and AUC score of 92.9%.

V. Conclusion

In this article, we proposed a model that could be used during the multidisciplinary consultation meeting (MCM), as a personalized prediction tool for the systematic management of patients with early breast cancer. Our model predicts the risk of metastatic relapse after four years for breast cancer patients likely to receive adjuvant therapy. This prediction can help decision-making in order to improve therapeutic management and increase the overall survival and quality of life of patients. We also presented a comparative study on the efficiency and effectiveness of the SVM, NB, k-NN and LR algorithms in terms of accuracy, precision, and sensitivity to find the best classification precision. The results obtained show that SVM has proven its efficiency and achieves the best performance in terms of precision and low error rate.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[2] Q. W. Lopez, "Évaluation de la réponse aux traitements et détermination de facteurs prédictifs et pronostiques dans le cancer du sein luminal (récepteurs hormonaux positifs/HER2-)," p. 222.

[3] E. Deluche and J.-Y. Pierga, "Chimiothérapie et femme jeune dans le cancer du sein : quelle prise en charge ?," Bulletin du Cancer, vol. 106, no. 12, pp. S19–S23, Dec. 2019, doi: 10.1016/S0007-4551(20)30043-6.

[4] A. Mailliez, C. Decanter, and J. Bonneterre, "Chimiothérapie adjuvante de cancer du sein et fertilité : estimation de l'impact, options de préservation et place de l'oncologue," Bulletin du Cancer, vol. 98, no. 7, pp. 741–751, Jul. 2011, doi: 10.1684/bdc.2011.1391.

[5] L. Hood and M. Flores, "A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory," New Biotechnology, vol. 29, no. 6, Art. no. 6, Sep. 2012, doi: 10.1016/j.nbt.2012.03.004.

[6] R. Snyderman, "Personalized health care: From theory to practice," Biotechnology Journal, vol. 7, no. 8, Art. no. 8, Aug. 2012, doi: 10.1002/biot.201100297.

[7] I. Greenwalt, N. Zaza, S. Das, and B. D. Li, "Precision Medicine and Targeted Therapies in Breast Cancer," Surgical Oncology Clinics of North America, vol. 29, no. 1, Art. no. 1, Jan. 2020, doi: 10.1016/j.soc.2019.08.004.

[8] J. C. O'Donnell, "Personalized Medicine and the Role of Health Economics and Outcomes Research: Issues, Applications, Emerging Trends, and Future Research," Value in Health, vol. 16, no. 6, Art. no. 6, Sep. 2013, doi: 10.1016/j.jval.2013.06.004.

[9] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[10] P. Jayapaul, A. Balasundaram, K. P. D. Seturamalingam, and K. Sekar, "Performance analysis of machine learning techniques for the prediction of breast cancer in big data environment," Erode, India, 2020, p. 140006. doi: 10.1063/5.0011116.

[11] M. M. Ibrahim, D. Ahmed, and R. Ahmed, "Deep Learning Hybrid with Binary Dragonfly Feature Selection for the Wisconsin Breast Cancer Dataset," IJACSA, vol. 12, no. 3, 2021, doi: 10.14569/IJACSA.2021.0120314.

[12] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, "Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data," IJACSA, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110808.

[13] T. A. Khan, K. A., S. Nasim, M. Alam, Z. Shahid, and M. S. Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease Classification," IJACSA, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111170.

[14] G. Battineni, N. Chintalapudi, and F. Amenta, "Performance analysis of different machine learning algorithms in breast cancer predictions," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 6, no. 23, p. 166010, Sep. 2020, doi: 10.4108/eai.28-5-2020.166010.

[15] W. Kim, K. S. Kim, and R. W. Park, "Nomogram of Naive Bayesian Model for Recurrence Prediction of Breast Cancer," Healthc Inform Res, vol. 22, no. 2, p. 89, 2016, doi: 10.4258/hir.2016.22.2.89.

[16] "Performance of Support Vector Machine Kernels (SVM-K) on Breast Cancer (BC) Dataset," ijrte, vol. 8, no. 2S7, pp. 412–417, Sep. 2019, doi: 10.35940/ijrte.B1076.0782S719.

[17] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," Cancer Inform, vol. 2, p. 117693510600200, Jan. 2006, doi: 10.1177/117693510600200030.

[18] W. Kim et al., "Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine," J Breast Cancer, vol. 15, no. 2, p. 230, 2012, doi: 10.4048/jbc.2012.15.2.230.

[19] S. Mook et al., "Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study," The Lancet Oncology, vol. 10, no. 11, pp. 1070–1076, Nov. 2009, doi: 10.1016/S1470-2045(09)70254-2.

[20] G. C. Wishart et al., "RPeRseaErcDh aIrCticlTe : a new UK prognostic model that predicts survival following surgery for invasive breast cancer," Breast Cancer Research, p. 10, 2010.

[21] M. Fekih et al., "Utilisation de référentiels et hétérogénéité décisionnelle des indications de chimiothérapie adjuvante dans les cancers du sein exprimant les récepteurs hormonaux, HER2-négatifs: résultats d'un sondage national en France," Bulletin du Cancer, vol. 101, no. 10, pp. 918–924, Nov. 2014, doi: 10.1684/bdc.2014.2030.

[22] Y.-J. Tseng et al., "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies," International Journal of Medical Informatics, vol. 128, pp. 79–86, Aug. 2019, doi: 10.1016/j.ijmedinf.2019.05.003.

[23] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," Clinical Epidemiology and Global Health, vol. 7, no. 3, Art. no. 3, Sep. 2019, doi: 10.1016/j.cegh.2018.10.003.

[24] P. Chapman et al., "Step-by-step data mining guide," p. 76.

[25] H. L. Afshar, M. Ahmadi, M. Roudbari, and F. Sadoughi, "Prediction of Breast Cancer Survival Through Knowledge Discovery in Databases," Global Journal of Health Science, vol. 7, no. 4, p. 7, 2015.

[26] A. Gonçalves, J. Moretta, F. Eisinger, and F. Bertucci, "Médecine personnalisée et cancer du sein : médecine anticipatoire, évaluation pronostique et ciblage thérapeutique," Bulletin du Cancer, vol. 100, no. 12, Art. no. 12, Dec. 2013, doi: 10.1684/bdc.2013.1856.

[27] C. Georges-Tarragano, F. Tapié de Celeyran, J. Platon, and J.-L. Misset, "Décider en cancérologie dans les situations médicosociales complexes: les réunions de concertation pluriprofessionnelles médicosociales et éthiques à l'hôpital Saint-Louis de Paris," Oncologie, vol. 16, no. 1, pp. 55–62, Jan. 2014, doi: 10.1007/s10269-014-2368-5.

[28] M. Scimeca et al., "Novel insights into breast cancer progression and metastasis: A multidisciplinary opportunity to transition from biology to clinical oncology," Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, vol. 1872, no. 1, pp. 138–148, Aug. 2019, doi: 10.1016/j.bbcan.2019.07.002.

[29] H. Rizki, C. Hillyar, O. Abbassi, and S. Miles-Dua, "The Utility of Oncotype DX for Adjuvant Chemotherapy Treatment Decisions in Estrogen Receptor-positive, Human Epidermal Growth Factor Receptor 2-negative, Node-negative Breast Cancer," Cureus, Mar. 2020, doi: 10.7759/cureus.7269.

[30] F. Penault-Llorca and N. Radosevic-Robin, "Ki67 assessment in breast cancer: an update," Pathology, vol. 49, no. 2, Art. no. 2, Feb. 2017, doi: 10.1016/j.pathol.2016.11.006.

[31] C. Criscitiello et al., "High Ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in Luminal B HER2 negative and node-positive breast cancer," The Breast, vol. 23, no. 1, pp. 69–75, Feb. 2014, doi: 10.1016/j.breast.2013.11.007.

[32] P. Cerda, "Similarity encoding for learning with dirty categorical variables," Mach Learn, p. 18, 2018.

[33] M. N. Murty and V. S. Devi, Pattern recognition: an algorithmic approach. London: Springer, 2012.

[34] G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., An introduction to statistical learning: with applications in R. New York: Springer, 2013.

[35] F. C. Pampel, Logistic regression: a primer. Thousand Oaks, Calif: Sage Publications, 2000.

[36] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

Effective ANN Model based on Neuro-Evolution Mechanism for Realistic Software Estimates in the Early Phase of Software Development

Ravi Kumar B N[1]
Dept. of Computer Science and Engineering
BMS Institute of Technology and Management
Bangalore, India

Dr. Yeresime Suresh[2]
Dept. of Computer Science and Engineering
Ballari Institute of Technology and Management
Ballari, India

***Abstract*—There is no doubt that the software industry is one of the fastest-growing sectors on the planet today. As the cost of the entire development process continues to rise, an effective mechanism is needed to estimate the required development cost to control better the cost overrun problem and make the final software product more competitive. However, in the early stages of planning, the project managers have difficulty estimating the realistic value of the effort and cost required to execute development activities. Software evaluation prior to development can minimize risk and upsurge project success rates. Many techniques have been suggested and employed for cost estimation. However, computations based on several of these techniques show that the estimation of development effort and cost vary, which may cause problems for software industries in allocating overall resources costs. The proposed research study proposes the artificial neural network (ANN) based Neural-Evolution technique to provide more realistic software estimates in the early stages of development. The proposed model uses the advantages of the topology augmentation using an evolutionary algorithm to automate and achieve optimality in ANN construction and training. Based on the results and performance analysis, it is observed that software effort prediction using the proposed approach is more accurate and better than other existing approaches.**

***Keywords*—Software cost estimation; COCOMO-II; neuro-evolution; artificial neural network; genetic algorithm**

I. INTRODUCTION

The software industry is undoubtedly one of the greatest innovations in the modern world [1]. The software development process broadly requires various discrete actions such as understanding the client requirements, analysis, preparing the user requirement specification, technical requirement specification, software requirement specification, and hardware requirement specification in the initial stages [2]. Further actions architecture design of the software, design of the modules, coding, integration, testing, and debugging. The overall development cost estimation depends on the individual cost and efforts required for each of the actions involved in the SDP. However, estimating the cost in software development has been a challenge facing researchers and professionals in software engineering over the past few years. The purpose of cost estimation is to help with decisions made during the development of a software project. Many factors affect the accuracy of cost estimation. If the cost is underestimated, the project may be delayed, lack implemented features, or not be completed. On the other hand, an overestimated cost can lead to higher software costs, a waste of resources, and even loss of opportunities for competing markets [3]. These factors can have negative consequences for the project, the development organization, and the customers. Thus, the quality of estimates can affect the quality of the software project.

Many software cost estimation models have been developed and improved, which can be categorized into algorithmic and non-algorithmic models [4]. In algorithmic cost model (ACM), typically a mathematical model or expressions are formulated using factors like i) source line of codes (SLOC), ii) risk calculation, and iii) skill levels obtained from the historical records; however, it fails to enumerate many vital factors including i) complexities, ii) reliability and experiences of the projects and due to this, it leads to the imprecise estimation. The constructive cost model- COCOMO is the most popular method in this category [5]. Further, it has evolved as COCOMO-II and has been widely used to design software cost predictors with various strategies considering basic cost indicators like lines of codes (LOC) and the function points [6-7]. The non-algorithmic approach is basically concerned with soft-computing approaches that overcome the limitations of the algorithmic model. The soft-computing approaches handle a better approximation of the solutions of the complex problems where many nonlinear and uncertain parameters are involved. Table I highlights the comparison of algorithmic and non-algorithmic models. Specifically, the existing approaches for the estimation, such as COCOMO and iii) function point-based model, all lack providing desirable accuracy as they ignore many of the critical drivers. So, these methods limit their applicability in the real-time scenario. In order to address these challenges, the soft-computing approaches are being extensively attracted the focus of the researchers by including approaches either individual or by hybrid techniques like- swarm optimization, fuzzy logic, genetic algorithm, machine learning, and neural network [8-10]. The advantage of the soft-computing approach is that it approximates the solutions created by the mess due to nonlinear factors that are uncertain and imprecise. In recent years, neural networks have gained prominence in software development. However, the literature presents several studies on applying neural networks and machine learning techniques

to estimate cost [11-12]. However, there is no consensus on which method best predicts software costs. The neural network architecture involves different configuration and hyperparameters such as layers, neuron nodes, transfer function, and learning parameters (weights and biases). Generally, the design of the learning model is specific to the particular data set and problem context. If the same model is introduced with a different dataset, it may not perform similarly. Therefore, the parameters mentioned above affect network performance. However, the evolution of models that produce good results in different environments is still a driving force for current research work. This paper suggests a unique approach to software development cost estimation based on Neuro-evolution. The proposed Neuro-evolution approach implements a mechanism of artificial intelligence (AI) that employs an evolutionary algorithm to generate optimal Artificial Neural Network (ANN) architecture. Further, the constructed ANN model in the proposed work is trained to adopt characteristics of software attributes using the previous dataset to produce accurate software estimates.

TABLE I. ANALYSIS OF ALGORITHMIC AND NON-ALGORITHMIC TECHNIQUES

Techniques	***Category***	***Advantages***	***Limitations***
Analogy	Non-Algorithmic	Independent of new resources	Dependent on past information & huge data requirement.
Expert-based		Highly responsive and fast process	Biased outcome
Bottom-Up		Stable	Inaccurate timings & needs huge data
Top-Down		Faster & low cost	less stable outcome & decisions
COCOMO	Algorithmic	Flexible analysis, input modification, & clear outcomes	Inaccurate estimates & practically infeasible
Function Point		Tool independent	Not good enough
Neural Network	Machine learning	Precise predictive estimates	Highly dependent on the dataset and no standard rule for implementation

The ANN model constructed is a feedforward neural network utilizing backpropagation learning mechanisms. The entire configuration and learning parameter is realized with the evolutionary algorithm, particularly a genetic algorithm (GA) implemented via the Neuro-evolution concept. The proposed study aims to achieve:

- A unique ANN model with an optimal selection of its parameters, including the size of hidden layers, number of neuron units at each layer, and transfer functions, from the given interval (linear, Relu, and sigmoid).
- The stable training process of the constructed ANN model that supports large training data samples.
- Self-adjustment in the weight and biases in an optimal manner from the training samples.
- Enhanced generalization in the training phase and efficient identification of dependencies of the predicted values from the input observations.
- Higher accuracy in the prediction to achieve realistic estimates of the cost required for the software development compared to the existing techniques.

The remaining sections of this paper are organized in the following manner: Section-II presents the review of the literature in the context of software cost and effort estimations; Section III discusses the material and methodology adopted in the proposed work; Section IV presents the system design and implementation procedure adopted in the proposed system; Section V presents the outcome and discusses the performance of the proposed system concerning its scope and effectiveness compared to the existing approaches, and finally, the entire contribution of the proposed work is summarized in Section VI.

II. RELATED WORK

Currently, the literature consists of several types of techniques and schemes for software cost estimation and prediction. This section discusses some of the recent research works carried in the context of enhancing prediction of the cost required for software development.

A. Algorithmic Approaches

The algorithmic approaches are concerned with mathematical models or expressions for cost predictions. To date, various methods have been suggested based on the algorithmic approaches. Work carried out by Kumawat, and Sharma [13] focuses on estimating the size metric for computing the cost required for the software project development (SPD). The authors have used the function point analysis (FPA) technique to compute cost estimates. The work of Khan et al. [14] suggested a cost estimation model by customizing features of the COCOMO-II that integrates additional cost drivers for computing the estimates of actual cost and effort required for SDP. Similarly, the study of Keil et al. [15] has introduced a different version of COCOMO-II to fit in the context of global software development (GSD). Two additional cost drivers are added in this version of cost drivers concerning collaboration and communication among different sites. The researchers in the above-discussed literature have tried to provide a significant contribution. All the factors are determined and devised based on the literature analysis and researchers' knowledge. However, there is a lack of empirical support, effective benchmarking, and validation of the scope of the suggested schemes. The authors in the study of Menzies et al. [16] have introduced a tool that encompasses case studies and previous experience to reduce the execution time, the effort required, and the number of defects in the project's development. Their results were obtained from small data sets, and they recommend conducting other tests where large volumes of information are handled. They do not explicitly use control indicators from other areas of knowledge, for example, to measure human and logistical resources. In the existing literature, few extensions to COCOMO were suggested, including dynamic multistage models to meet the analytical needs of prototyping SPD models. These models consider the

dynamics of varying requirements, system design, and other strategies, but all lack desirable accuracy as they ignore many critical drivers. So, these methods limit their applicability with varied IDEs models, languages, and tools.

B. Non-Algorithmic Approaches

The non-algorithmic approach generally implies the soft-computing techniques that handle ambiguity and nonlinearity in the cost estimation techniques. The previous section discusses the conventional approaches regarding software cost and effort estimation. However, software project requirements constantly change over time, which also causes the estimates of cost and effort to change. The researchers realized the need for soft computing approaches that include machine learning techniques, fuzzy logic, and various metaheuristic method. This section discusses the existing soft computing approaches for software effort and cost estimation to analyze the current research trend. Nandal and Sangwan [17] a hybrid Bat and Grautational algorithm is used to estimate the effort of software, whereas fuzzy regression models are used to overcome the problem of imprecise in the dataset for the prediction software effort (Nassif et al. [18]). All these approaches provide a good solution but at the cost of huge computational complexity. The application of evolutionary algorithms like GA is used in the study of Zaidi et al. [19] and Reena et al. [20] to optimize the coefficients of different estimation models in the presence of nonlinear data. The approach of intelligent techniques like the neural network deals with the complexities and uncertainty in the software effort estimation is presented in Venkataiah et al. [21] [22]. Few recent research studies have also focused on applying the hybrid approach in the SPD process. The joint approach of nature-inspired algorithm and ML is adopted by authors in [23-25] to compute the estimates of effort in project development. The work of Singh et al. [26] evaluated different ML techniques in the software effort estimation. The outcome reported in this study showed better performance achieved by LR in terms of error percentage analysis. A neural network approach [27-28] has also been widely accepted in software cost estimation. In the work of Choetkiertikul et al. [29], a long short-term memory (LSTM) and recurrent highway network (RHN) are employed to estimate the effort required for completing user stories or issues. Also, Bayesian Network is used to estimate the work time required in the SPD process [30].

C. Motivation of the Research

A wide range of schemes and techniques have been described in the literature for predicting SPD's costs. The recent literature has been observed more focused on applying metaheuristic techniques, neural networks, and machine learning algorithms. Building a model based on the dataset is difficult due to the complexity and nonlinearity involved in the data attributes. Also, the learning model's design is affected by a variety of factors concerned with network parameters, data modeling, and feature engineering. Apart from this, the factors that determine the connectivity among nodes are complicated to analyze before the training phase to develop an ideal network. Generally, the building and training of the learning model involves a lot of human effort and is specific to the particular context, which is a significant concern as software attributes vary over time. However, even small changes in parameters can dramatically alter the result of the trained model.

A unique model with accurate estimation is presented based on the neuro-evaluation augmenting topology to evolve with an optimized ANN architecture to address and overcome these problems. This type of approach for the cost estimation problem has not yet been applied to the software cost estimation problem. The proposed study aims to explore the effectiveness of augmenting the topology mechanism to automate the construction and training of the ANN model that generates better solutions.

III. Materials and Methodology

The material used for evaluating the proposed model is the COCOMO dataset. The methodology used for designing and developing the proposed ANN model for cost estimation is based on the Neuro-evolution AI technique, which constructs an optimal ANN model using a genetic algorithm. This section briefly highlights the dataset and methodology adopted in the proposed system.

A. Dataset

The COCOMO (Constructive Cost Model) is a widely known software estimation model introduced by Barry Boehm [31]. This model utilizes an approach of statistical correlation between software attributes and lines of the code. In other words, it basically adopts regression analysis with the responsible parameters that are representative of the estimates of the cost required in software development. In the current research work, the study uses the COCOMO NASA-2 dataset publicly accessible at the promise software engineering repository. This dataset consists of a total of 24 vital cost attributes from 93NASA projects.

B. Artificial Neural Network

In recent years, ANN has received wide attention to address complex nonlinear problems in various fields such as computer vision, image processing, natural language processing, and many more. ANN can be viewed as a function approximator that takes an input from observation state and maps to the output state (decision), such that: $f(x) \rightarrow y$. Typically, the function approximators consist of neurons, often referred to as cells or units, composed of summation and activation functions. The typical function of ANN cell is described in Fig. 1 as follows:

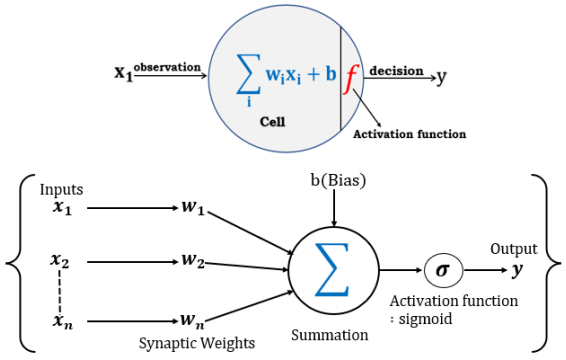


Fig. 1. Typical Function of ANN Cell.

In Fig. 1, the architecture of the basic ANN cell is described where x is the n input such that:$x \in [x_1, x_2, x_3 \cdots x_n]$, w indicates synaptic weight, such that: $w \in [w_1, w_2, w_3 \cdots w_n]$. Each weight 'w' are associated with input sample $'x'$ both together served as input to the cell function, where all x is multiplied with w and are summed with biased (b) using summation function as described as follows:

$$x \cdot w = (x_1 \times w_1) + (x_2 \times w_2) + \cdots + (x_n \times w_n) \quad (1)$$

Equation 25 describes the dot product of vector x and vector w, and their summation is given in equation 26 as follows:

$$\sum = x \cdot w \quad (2)$$

The weights'w_i' can be considered as a strength of the association between cells, and it also decides how much influence the given input will have on the cell's output. Another essential component of the ANN cell is the offset value added to the summation of dot product $x \cdot w$. This offset value is often called a bias that allows shifting the phenomenon of the nonlinear activation function to produce the expected result correctly to the output state. Moreover, the w and b are also often called learning parameters of the ANN model; the relationship between w and b can be numerically represented as follows:

$$(x \cdot w) + b \quad (3)$$

Equation 3 is then passed to the nonlinear function, which is generally a sigmoid function that enables nonlinearity in the ANN cell as numerically represented as follows:

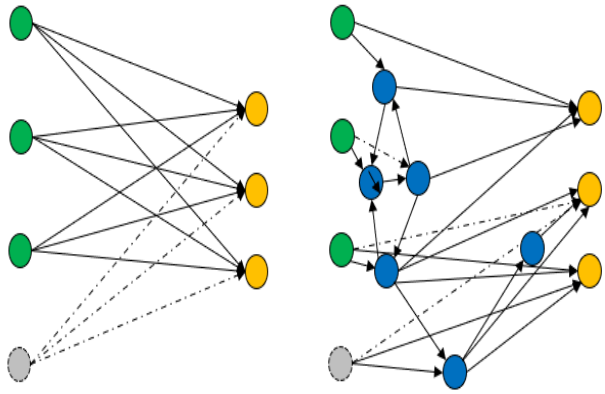
$$y = \sigma(x \cdot w) + b \quad (4)$$

Where y denotes the output of the cell and nonlinear σ sigmoid function. Sigmoid or Logistic: takes a real-valued input and returns output in the range [0,1]. The ANN cells are arranged into several layers, typically classified as input layers, hidden layers and output layers all interconnected to each other.

Usually, the topological structure of the artificial neural network is selected based on empirical analysis, and the learning parameters are determined using the training process, which is related to the trial-and-error process. Therefore, developing an ANN model is not a big problem. However, training ANN models to accomplish certain tasks is a real challenge. In this regard, Neuro-Evolution can be an effective mechanism for determining the optimal topology of neural networks and learning parameters (weights and biases) to construct an ideal ANN model.

C. Neuro-Evolution of Augmenting Topologies

Neuro-Evolution of Augmenting Topology (NEAT) is a neuroevolutionary AI technology that deals with topology augmentation to automate the construction and training of ANN models using evolutionary algorithms (EA) [32]. The EA in NEAT is a kind of genetic algorithm (selection, crossover, and mutation), which allows the evolution of ANN units, learning parameters (weight and biases), and structure, trying to determine stability between the fitness of the obtained solution and assortment. Fig. 2 shows a sample visualization of the topology construction of ANN using the NEAT algorithm.



(a) Initial Architecture of AN (b) Augmented Topologies of ANN.

Fig. 2. Topology Construction of ANN using NEAT.

In the above Fig. 2, visualization of initial topology (a) and final topology construction of ANN model (b) after several iterations is shown using NEAT. The flow process of topology augmentation in the construction and training of the ANN model is shown in Fig. 3.

The mechanism of topology augmentation for the optimal ANN model requires the initialization of variables concerning network hyperparameter and loss function. The initialization of hyperparameter variables (such as learning rate and the number of neurons) is crucial to determine the training performance of the network during the crossover and mutation process of EA. On the other hand, the loss function determines the optimality of the neuron genes (bias) and synapse genes (weight) in the learning phase. The loss function in NEAT is also regarded as a fitness function, and a set of neuron genes and synapse genes are called genomes.

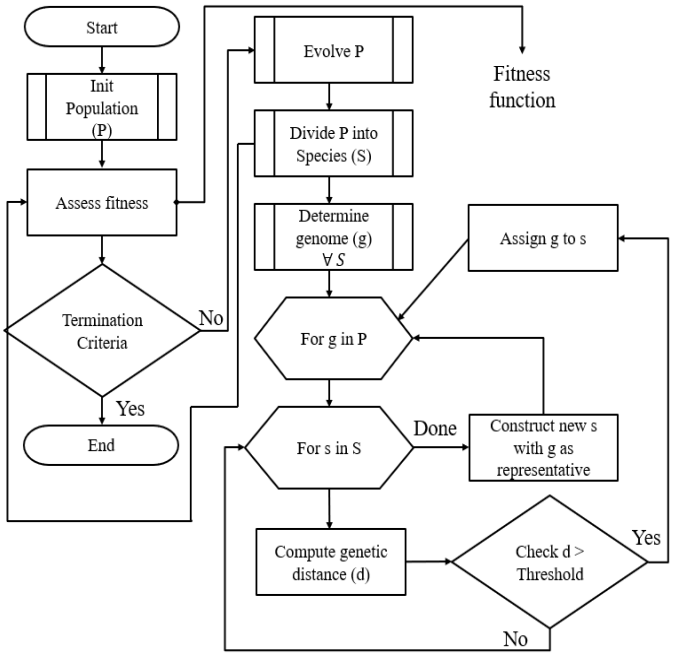


Fig. 3. Flow Process of Topology Augmentation using NEAT.

The algorithm generates a genome considering single input and output layer during the initialization of an initial set of solution candidates (population). Therefore, in the first generation, the genomes only vary in weights and biases but not network topology. After assessing the fitness value of each genome, the algorithm stops if the termination criterion is met. Otherwise, it generates a new set of solution candidates by executing crossovers phase y between genomes and then performs mutations in the subsequent offspring. All these processes are carried out randomly, and prior to computing the fitness of neuron genes and synapse genes, i.e., optimality of weight and biases, the algorithm splits the set of solution candidates into species (a particular class with the common characteristics) based on the computation of the genetic distance between each set of neuron weight and biases. The computation of the genetic distance is carried out using the following numerical equation:

$$d = d_b + d_w \tag{5}$$

The above equation 6 represents the computation of distance (d) based on the summation of neuron (d_b i.e., bias) and synapse (d_w i.e., weight). The computation of the d_b and d_w are shown in equations 6 and 7 as follows:

$$d_b = c_n \times \frac{\Delta_b}{\max(B(g_1),B(g_2))} \tag{6}$$

$$d_w = c_s \times \frac{\Delta_w}{\max(W(g_1),W(g_2))} \tag{7}$$

Where c_n and c_s are the user-defined variables for fine-tuning the model parameters.

IV. Proposed Cost Estimation Model

This section discusses the proposed cost estimation implementation procedure based on the ANN model determined using the NEAT algorithm discussed in the previous section. In the proposed study, the cost estimation problem is being studied as a regression problem rather than an optimization problem to predict kilo line of code (KLOC). The proposed cost estimation model design involves three core modules; namely, i) data exploration module ii) data preprocessing, and iii) design of ANN Model.

A. Dataset Exploration

In the current study, the data is available on the NASA website. The data is downloaded by sending an HTTP GET request to the respective URLs. When the request is sent, the data can be retrieved in the form of an a.arff file. However, this is not readable readily by our system. Hence, the data is sub-set from the 'Arff file', which contains 10 parts, including {Title, Past Usage, Relevant Information, Number of instances, Number of attributes, Attribute information, Missing attributes, Class distribution, Data}. The sub-set extracts only the Data. The Data Store stores the data in the form of a simple CSV file. Each column is separated by a (delimiter), and a new line character separates each sample. Many data science platforms can read and process this format, including pandas used in the current study. The data imported into the numerical computing environment (NCE) describes 124 entries ranging from the index number 0 to 123 with 24 columns. The dataset consists of 24 variables with type numeric and two categorical variables. The memory taken to upload the data is more than 25 KB. Table II presents a statistical description of all the 25 predictors and an output KLOC. The closer shows that the counts of all the parameters are identical to the number of samples, which indicates there are no missing values. The differential between the consecutive pair between {0, 25%}, {25%, 50%}, {50%, 75%} and {75%, 100%} sometimes are not less than standard deviation (σ) that means there is the presence of outliers in the data, as well if RMSE and MAE of the model have a difference more than mean KLOC then outliers need to be corrected. Another important observation on the dataset is that certain parameters show a specific correlation with the effort. The correlations are either negative correlation or positive correlation. In positively correlated parameters, the effort decreases with a decrease in the parameter's values, whereas, in negatively correlated parameters, the effort decreases if the parameters increase. The positively correlated parameters are the cost drivers (CD) ∈ {acap, pcap}, and negatively correlated parameters such that CD ∈ {rely, Cplx, data, time, stor, sced}. Further, on the analysis of co-efficient using linear regression analysis, it is found that reduced reusability (ruse) and 'site' have a higher multiplier effect on cost/effort compared to other CDs, as evident in Fig. 4. It is clear that the correlation of data points with the actual effort is highly non-uniform in nature. Therefore, a custom feature engineering process for the proposed ANN-based CEM is being carried out.

B. Preprocessing

In this section, the preprocessing operation is carried out from the perspective of the feature engineering task and the extraction of suitable input for the proposed learning model. The core module in this stage contains i) correlation analysis and ii) dataset normalization. In the correlation analysis, the relationships between various variables are analyzed using a mathematical approach that helps find correlations between various cost drivers. The formula for correlation is shown in the equation as follows:

$$r_{x,y} = \frac{\sum(x_i - x').(y_i - y')}{\sqrt{\sum(x_i - x')^2.\sum(y_i - y')^2}} \tag{8}$$

Where, x_i and y_i denotes cost drivers, x' and y' are means values of the cost drivers and $r_{x,y}$ is the correlation factor between x and y that ranges from -1 to +1. As it can be observed from the formula if $x \propto y$, which means that $x = ky$, then the following outcome is achieved when the same is substituted in equation 9.

$$r_{x,y} = \frac{\sum(x_i - x').k(x_i - x')}{\sqrt{\sum(x_i - x')^2.k^2\sum(x_i - x')^2}} \tag{9}$$

$$r_{x,y} = \frac{\sum k(x_i - x')^2}{k\sqrt{(\sum(x_i - x')^2)^2}} \tag{10}$$

$$r_{x,y} = \frac{k\sum(x_i - x')^2}{k\sum(x_i - x')^2} \tag{11}$$

$$r_{x,y} = 1 \tag{12}$$

The above equation 12 proves that when the two cost drivers are proportional, the correlation between them is one. Similarly, when one cost driver reduces and another cost driver

increases, in other words, x=k-ly, then the correlation is said to be -1 and considered as an ideal scenario when there is a perfect linear relationship between two CDs. However, a zero correlation refers to total randomness and no relation between two CDs. The correlation plot for among CDs is given in Fig. 5. It can be analyzed that there is a strong correlation between the 'prec', 'flex', 'resl' and 'team'. As it can be observed that except for exponential CDs such that {'prec', 'flex, 'resl', 'team' and 'pmat'} all other CDs have (>10%) correlation. Hence, all variable turns out to be significant while building an ANN model.

TABLE II. DESCRIPTIVE STATISTICS

Cost Drivers	count	mean	std	min	25%	50%	75%	max
ACT_EFFORT	124.0	563.334677	1029.227941	6.00	71.50	239.500	581.750	8211.00
prec	124.0	3.110000	1.292409	0.00	2.48	2.480	4.9600	4.960000
flex	124.0	2.618952	1.041618	0.00	103	2.030	4.0500	5.070000
resl	124.0	3.688871	1.403707	0.00	2.83	2.830	5.6500	6.010000
team	124.0	1.837097	1.094185	0.00	1.10	1.100	3.2900	4.660000
pmat	124.0	5.602984	1.288265	2.84	4.68	4.680	6.2400	7.800000
relay	124.0	1.078522	0.103427	0.85	1.00	1.100	1.1000	1.740000
Cplx	124.0	1.189892	0.163256	0.87	1.17	1.170	1.2125	1.740000
Data	124.0	1.014919	0.117179	0.90	0.90	1.000	1.1400	1.280000
Ruse	124.0	0.996935	0.014605	0.95	1.00	1.000	1.0000	1.070000
Time	124.0	1.124516	0.184476	1.00	1.00	1.000	1.2900	1.630000
Stor	124.0	1.107097	0.163149	1.00	1.00	1.000	1.1700	1.460000
Pvol	124.0	0.927406	0.095456	0.87	0.87	0.870	1.0000	1.150000
Acap	124.0	0.880276	0.101079	0.71	0.85	0.850	1.0000	1.016667
Pcap	124.0	0.918817	0.085625	0.76	0.88	0.895	1.0000	1.000000
pcon	124.0	1.000544	0.035766	0.81	1.00	1.000	1.0000	1.205000
Apex	124.0	0.925712	0.083496	0.81	0.88	0.880	1.0000	1.220000
Plex	124.0	1.004590	0.080974	0.91	0.91	1.000	1.0000	1.190000
ltex	124.0	0.966781	0.089415	0.91	0.91	0.910	1.0000	1.200000
Tool	124.0	1.115847	0.078542	0.83	1.09	1.170	1.1700	1.170000
Sced	124.0	1.043065	0.063760	1.00	1.00	1.000	1.1400	1.140000
Site	124.0	0.925040	0.017623	0.86	0.93	0.930	0.9300	0.947500
docu	124.0	1.024940	0.057830	0.91	1.00	1.000	1.1100	1.230000
Physical Delivered KLOC	124.0	103.443901	141.455891	0.00	20.00	51.900	131.7500	980.000000

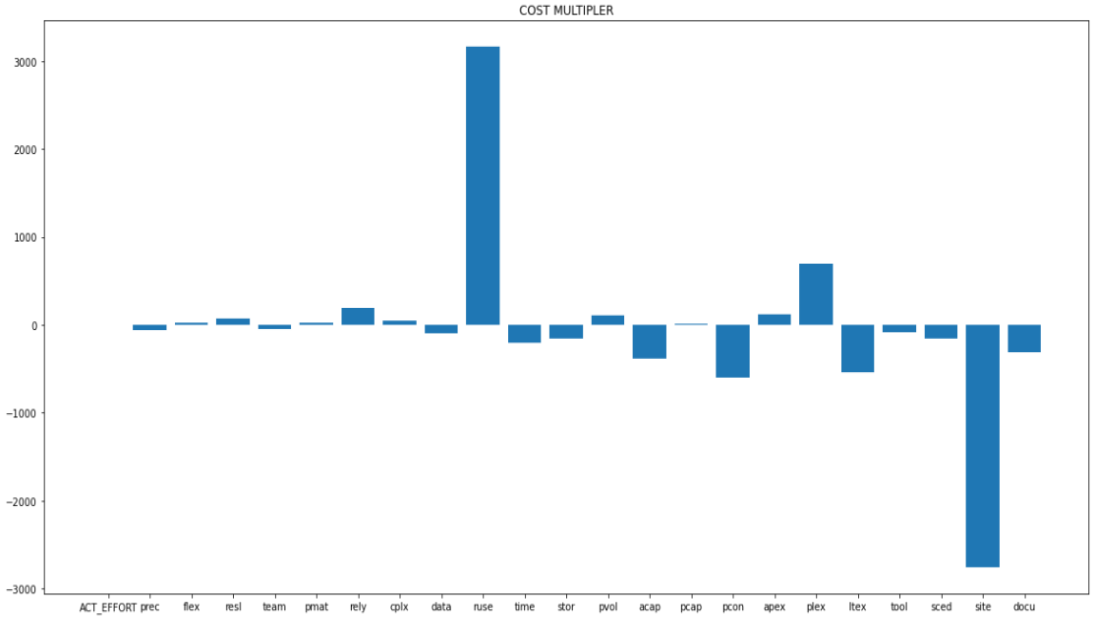


Fig. 4. Representation of Cost Multiplier.

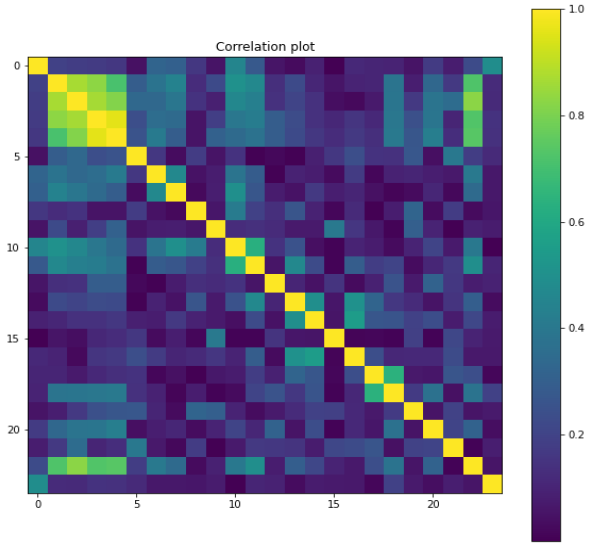


Fig. 5. Correlation Plot among CDs and KLOC.

In order to provide an input to a learning model, the input data is required to be in a vector form. Feature vectorization refers to converting a row of values into a usable vector. In this phase of implementation, the data is normalized with the help of the Min-Max scaling method. Further, each row is transposed and fed to neural networks. The typical formula for data normalization for feature vectors is numerically expressed in equation 13.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (13)$$

Where, x is the input data, i.e., original CDs feature samples, which is normalized using min and max function and rescaled in the range of [0,1], and x' normalized CDs feature samples which are further fed to the proposed learning model.

C. Design of the Proposed ANN Model

This section discusses the ANN model design and its implementation procedure with the support of the algorithmic steps. The implementation procedure utilizes the NEAT library of python executed in the Anaconda distribution. The dataset is split into training and testing sets, where 80% of the dataset is kept for the model training, and 20% of the dataset is kept for model testing. The design configuration of the proposed ANN model is carried out using neural evolution mechanisms, where the features from the input observation are considered for determining weights and biases. In this process, the optimality of the ANN architecture is determined through topology augmentation using a genetic algorithm. The configuration parameters considered in the ANN construction consist of hidden layers, neurons unit at each hidden layer, and a set of transfer functions. The proposed study considers three transfer functions: linear, Relu, and nonlinear sigmoid. On the other hand, mean square error (MSE) is considered a fitness function. Since the proposed study has considered MSE, the fitness evaluation is carried out based the less error. Therefore, the inverse roulette selection (IRS) technique is considered for the proportionate fitness selection. The core configuration and training process of ANN construction using topology augmentation is shown in Fig. 6. The topology augmentation begins with the initialization of population (a set of candidate solutions), basically a pool of random neural networks. The process iterates several times, which is also called a generation where the algorithm chooses the optimal ANN based on the fitness value, which is then further cross overed according to the selection/decision process.

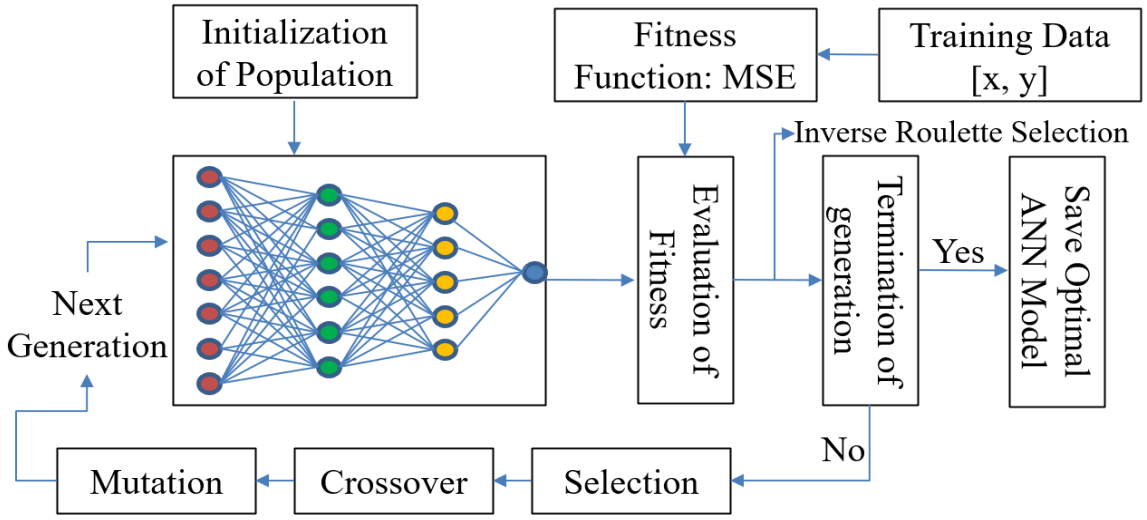


Fig. 6. Generation of Optimal ANN Model using Neuro-evolution Technique.

Afterward, a new ANN model is generated, and after mutation, an evolved version of the ANN model is further carried out for the training process. All these processes continue until the termination criteria are met. This termination criterion is based on the specified number of generations, wherein each generation, the trained model is evaluated and selected according to the prediction performance. The implementation steps for the above-discussed procedure are mentioned as follows:

Algorithm:1 Neuro evolution training

Step 1. Create population pool

In this step, the population pool is generated, a set of random neural networks with random layers and neurons and random activation functions. Inputs to the algorithm are given in the form of a finite number of layers and neurons and, at the same time, a set of activation functions. The activation functions allowed are, Sigmoid, Linear, and relu.

Step 2. Evaluate fitness of the population

The MSE fitness function measures the fitness of the population. The MSE of the input data is considered with the output in the training set.

Step 3. Select the fittest individual to reproduce

The inverse Russian roulette process selects the individuals for the repopulation pool. The lower the fitness function value, the higher the probability of the selection. The following equation decides the probability of selection.

$$P_i = 1 - \frac{MSE_i}{\sum_{i=1}^{n} MSE_i} \quad (14)$$

Step 4. Repopulate using copies of the fittest network

Most fit individuals among the population are selected and used for further processing. The crossover of these individuals is made here, and also mutation is applied according to the mutation probability.

Step 5. Introduce normally distributed mutations to the network weights

The neural networks are finalized in this step, and the newly formed networks are introduced to the population pool.

V. Result and Performance Analysis

This section discusses the performance metrics followed by outcome analysis to justify the scope and effectiveness of the proposed system.

A. Neuro Evolution Model Parameters

The design and development of the proposed system are done using python programming language and execution on Anaconda. The parameters considered for executing proposed neuroevolutionary technique for obtaining optimal ANN model is mentioned in Table III.

The parameter namely population size is the total number of offspring (networks) present in each generation and total number of generations is number of times the fitness is measured. In 15% of the cases a new neuron is added to the network. In 10% of the cases an existing neuron is deleted from the network. Addition and deletion of neurons happen within a single generation. Either relu, sigmoid or linear activation functions are chosen. Initial bias is assigned according to the normal distribution. Maximum value of weights and bias are set to 30 however the minimum weight is set to 0 in order avoid negative values. At the same time, minimum bias is set to -5 in order to cancel out certain values.

Mutation probability is 5%. This is necessary to display the stochastic nature of the system. After successful execution of the neuro-evolution training, the proposed algorithm returns optimal ANN model discussed in Table IV.

The architecture of the obtained ANN model is shown in Fig. 7. After evolution through several iteration, the neuro-evolution algorithm provides optimal number of layers and number of neurons unit at each layer as mentioned in Table IV.

TABLE III. Neuro-evolution Hyperparameters

Parameters	Values
Population size	200
Number of generations	100
Probability of adding a new neuron	0.15
Probability of deleting a neuron	0.1
Activation function	Sigmoid, Relu, Linear
Initial bias	according to normal distribution
Mutation probability	0.5
Minimum neuron bias	-5
Maximum neuron bias	30
Minimum weight	0
Maximum weight	30
Weight mutation probability	0.5

TABLE IV. Configuration Description of Obtained Optimal ANN Model

Layer	Number of neurons	Trainable parameters
Layer 1 (input)	24	N/A
Layer 2	10	(24*10) + 10 = 250
Layer 3	5	(10 * 5) + 5 = 55
Layer 4 (output)	1	(5 * 1) + 1 = 7
Loss Function (MSE)	-	-
Activation Function (Relu)	-	-
	Total neurons: 40	Total trainable parameters: 312

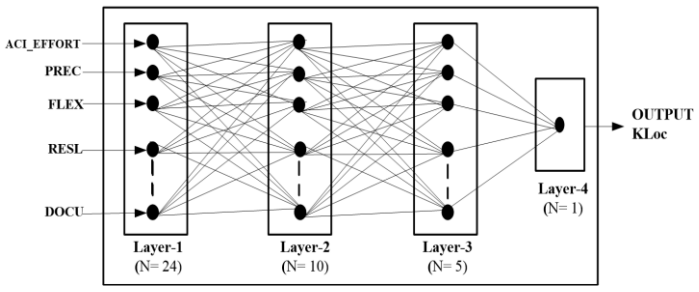


Fig. 7. Architecture of Optimal ANN Model.

B. Performance Metrics

1) MMRE (Mean Magnitude of Relative Error): The MMRE performance metric is the most common basis for the assessment of the effort estimation process. The matric MMRE is computed for the given dataset of software projects whose estimated efforts are compared with their actual efforts. The estimation process with minimum MMRE is considered to be the most accurate. The formula for calculating MMRE is given as Eq. 15.

$$\text{MMRE} = \frac{1}{\text{N}} \cdot \sum_{\text{i=1}}^{\text{n}} \frac{\lfloor (\text{y} - \text{y}\prime) \rfloor}{\text{y}} \quad (15)$$

Where, y is the actual effort, and y' denotes estimated work effort for project pi, and *N* is the total project (PI) under consideration. Mathematically, MMRE gives an average percentage of error between y and y'.

2) MSE (Mean Squared Error): MSE is being calculated in proposed implementations to analyze the performance of proposed methods over other LR and SVR. MSE is more critical function while building better models while optimizing the learning model. The formula for calculating MSE is given as Eq. 16.

$$\text{MSE} = \frac{1}{\text{n}} \sum_{\text{i=1}}^{\text{n}} (\text{y} - \text{y}')^2 \quad (16)$$

Where y is the actual effort, and y' denotes estimated work effort for project pi, and *N* is the total number of the project under consideration.

3) RMSE (Root Mean Square Error): Since the unit of MSE is squared, RMSE is the square root of MSE used since the unit of MSE is Nl^2 where Nl is the number of lines of code in the project. Though MSE is significant for optimizing the model, it would make no sense to human beings. Hence, the study considers $\text{RMSE} = \sqrt{\text{MSE}}$. Since the unit of RMSE is Nl, it can be assumed that the most probable range for y can be $\text{y} = \text{y}' \pm \text{RMSE}$. The computation of RMSE can be numerically represented as follows in eq. 17:

$$\text{RMSE} = \sqrt{\frac{1}{\text{n}} \sum_{\text{i=1}}^{\text{n}} (\text{y} - \text{y}')^2} \quad (17)$$

4) MAE (Mean Absolute Error): This is similar to MMRE, representing average absolute error instead of providing average percentage error. In MAE abs function is used to remove the error from simple error, and the average is calculated. Due to this, some of the extreme points, like outliers, will provide less significance; hence this measure is less sensitive to outliers. MAE can be numerically represented as follows in eq. 18:

$$\text{MAE} = \frac{1}{\text{N}} \cdot \sum_{\text{i=1}}^{\text{n}} \lfloor (\text{y} - \text{y}') \rfloor \quad (18)$$

Since the unit of MAE and output (actual cost) is the same, MAE represents total cost overrun or underrun.

5) Pred: PRED is the de facto standard for cost model accuracy measurement. It is called the percentage of predictions falling within the K% of the actual known value. The formula for PRED calculation is shown in equation 19:

$$\text{PRED} = \frac{1}{\text{n}} \sum_{\text{i=1}}^{\text{n}} \left| \frac{\text{EstimationEffort} - \text{ActualEffort}}{\text{Actual Effort}} \right| \text{K \%} \quad (19)$$

Where k% is the percentage error between AE and EE, PRED represents the percentage of a number of projects whose cost overrun or underrun is below 25% in some researches 30%.

C. Outcome Analysis

This section discusses the outcome obtained for the proposed system based on the comparative analysis. The proposed study implements two machine learning algorithms for the comparative analysis such as Linear regression (LR) and supports vector regression (SVR). In order to compare ANN with LR and SVR, the performance metrics MSE, RMSE, and MAE are considered. To justify the scope of the proposed optimal ANN model, the study also considers performance analysis with similar existing approaches such as estimation technique based on fuzzy-genetic [33] and based Dolphin optimization technique [34], Bat optimization [34], and combined Dolphin-BAT [34], the performance metric PRED and MMRE is used. The quantitative outcome obtained for the proposed system and its comparison is shown in Table V.

As it can be observed in Table V, that LR, SVR is associated with 151% and 128% errors, respectively, which means the predicted/estimated value could be more than twice as big as the actual value; therefore, making LR and SVR unfit for real-world implementations. However, even the most basic benchmarked algorithms (GA) are giving 29.9% error which is below 30%, which is an acceptable cost overrun ratio for software projects in general. It is also far below 77%, which is the average cost overrun ratio of the NASA project from which the dataset is collected. The overall numerical outcome shows the proposed ANN's effectiveness regarding the cost overrun ratio. Performance analysis regarding MAE is shown in Table VI.

TABLE V. QUANTITATIVE OBSERVATION IN TERMS OF MMRE

Methods	Performance Metrics
LR	1.510457
SVR	1.281522
GA	0.299469
BAT	0.1698
DOLPHIN	0.1665
DOLPHIN-BAT	0.14576
ANN	0.113518

TABLE VI. QUANTITATIVE OBSERVATION IN TERMS OF MAE

Methods	Performance Metrics
LR	119.266357
SVR	81.872095
ANN	22.151230

The performance metric MAE is used to calculate the performance of proposed methods over other LR and SVR. Since the unit of MAE is in Nl, the MAE value 22.15 obtained for ANN represents a number of lines of codes in the projects that may vary by 22,151 lines in ANN. An average developer writes 250 lines of production code per week (40 hours of working per week). An extra 22151 lines represent 88 weeks of work (3520-man hours). Considering that an average developer in the USA earns approximately $34 per hour, the total cost overrun might come to $119,680. In the cases of LR and SVR, the cost overrun is quite more than ANN, which is impractical for real-time implementation? The performance of the learning models implemented in this study regarding MSE is shown in Table VII.

The metric MSE is being considered in proposed implementations to assess the performance of the proposed ANN over other LR and SVR. MSE represents the overall training of the algorithm as it is used for optimization. Even though the MSE does not directly represent the algorithm's performance, it does represent the quality and level of training given to the algorithm. Lower MSE represents higher knowledge of the algorithm. More trainable parameters can store more knowledge among them. The MSE score is higher in both LR and SVR as they contain fewer trainable parameters than ANN. The quantified outcome indicates that ANN is less associated with error compared to LR and SVR. Therefore, it can be concluded that SVR and LR are subjected issue of underfitting. The performance analysis in terms of RMSE is mentioned in Table VIII.

Similarly, the metric RMSE is considered to evaluate the training performance of the learning models. The RMSE also helps to understand the requirement re-training model by the preprocessing step. From the quantified outcome, the proposed ANN scored 39.33 % RMSE and 22.15% MAE from Table VI, i.e., a difference of 17.18 % compared to mean KLOC of all projects, i.e., 103.44. This indicates minor variation with 16%-17%, which is within the acceptable limit of 20 %. The performance analysis regarding PRED is shown in Table IX.

PRED represents the ratio of projects which has less than a threshold percentage of cost overrun. Hence, this performance measurement is more practical than the other metrics since it represents the number of projects that will fall below the acceptable cost overrun ratio. In most of the studies, the threshold is set to 30%. In this study, 25% of the threshold value is considered to perform comparative analysis. From Table IX, it can be observed that the proposed model ANN achieved a higher PRED value, i.e., 68.91, compared to other ML methods and existing approaches. Bat, Dolphin, hybrid Dolphin-Bat, and the proposed ANN are more practical to implement as they have PRED value much higher than GA. But among them, the proposed ANN method has the highest PRED value, which indicates its suitability and scope in the real-world system. The following analysis mentions the overall improvement (%) of ANN concerning MMRE in Fig. 8 and PRED in Fig. 9 over other implemented ML models and existing approaches.

TABLE VII. QUANTITATIVE OBSERVATION IN TERMS OF MSE

Methods	Performance Metrics
LR	42545.810081
SVR	29240.145478
ANN	1547.247493

TABLE VIII. QUANTITATIVE OBSERVATION IN TERMS OF RMSE

Methods	Performance Metrics
LR	206.266357
SVR	170.997501
ANN	39.335067

TABLE IX. QUANTITATIVE OBSERVATION IN TERMS OF PRED

Methods	Performance Metrics
LR	2.335234
SVR	5.297425
GA	11.66
BAT	61.66
DOLPHIN	61.66
DOLPHIN-BAT	66.66
ANN	68.91522

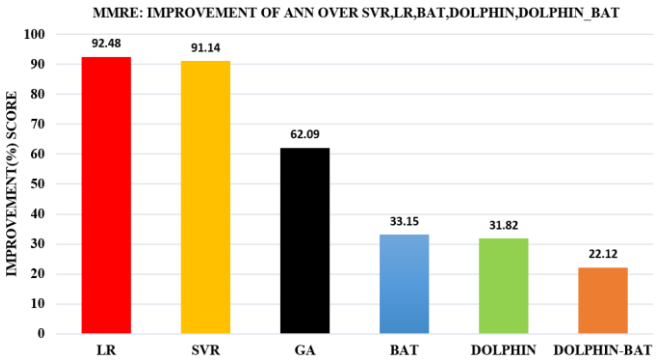


Fig. 8. MMRE Improvement (%) of ANN over SVR, LR and existing Methods.

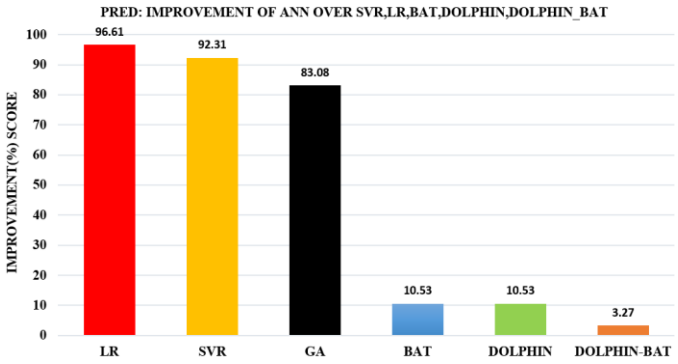


Fig. 9. PRED Improvements (%) of ANN over SVR, LR and Existing Methods.

The analysis from Fig. 8 shows ANN has achieved 92.4% improvement over LR, 91.14% improvement over SVR, and 62.09%, 33.15%, 31.82%, 22.12% over Fuzzy-GA, BAT, Dolphin, and Dolphin-Bat, respectively. The analysis from Fig. 9 shows that ANN has achieved 96.91% improvement over LR, 92.31% improvement over SVR, and 83.08%, 10.53%, 10.53%, 3.27% over Fuzzy-GA, BAT, Dolphin, and Dolphin-Bat, respectively. Hence, it can be seen that the proposed offers a good result regarding software cost estimates. The overall analysis shows effectiveness of the proposed neuro-evolution algorithm towards devising suitable learning model for achieving realistic estimates of the cost required in the initial stage of the software development process. Hence, the proposed research work suggested a technically-efficient method acquainted with recent trends and technologies to benefit real-world applications.

VI. Conclusion

The development of software projects involves various phases like initial planning, risk assessment, effort, and cost estimation. Among these, cost estimation is the key concern in the software industry. The conventional approaches do not provide accurate estimation due to the lack of precise system and cost drivers modeling. In this paper, the study has presented a novel and unique approach to predict realistic estimates of the cost needed to develop a software project. The proposed study applied a mechanism of neural evolution in conjunction with evolutionary technique, namely genetic algorithm top construct ANN, which predicts actual estimates of the cost required to develop a software. The application of neural evolution in ANN modeling proves its effectiveness and scope that it can compete with the existing techniques in terms of realistic estimates of the cost and effort. Once developed and trained, the proposed ANN can estimate the development costs in real-time as it computes cost estimates based on the responsible attributes required in the development of the software. The execution complexity grows linearly with the problem context and size of data samples. Based on the result analysis, it is observed that the proposed ANN is producing better results than other previously proposed algorithms and other machine learning models being implemented. The existing works adopted global optimization algorithms that require huge computing resources due to recursive operation in parallel. However, the proposed ANN model is constructed optimally using the mechanism of augmenting topology, and it better adopts generalization of the feature from the input observations, therefore, providing accurate estimates of the cost compared to the existing approaches.

References

[1] Alt, R., Leimeister, J.M., Priemuth, T. et al. Software-Defined Business. Bus Inf Syst Eng 62, 609–621 (2020).

[2] Trendowicz, A., 2013. Software Cost Estimation, Benchmarking, and Risk Assessment: The Software Decision-Makers' Guide to Predictable Software Development. Springer Science & Business Media.

[3] Mittas, N. and Angelis, L., 2013, September. Overestimation and underestimation of software cost models: Evaluation by visualization. In 2013 39th Euromicro Conference on Software Engineering and Advanced Applications (pp. 317-324). IEEE.

[4] Khan, B., Khan, W., Arshad, M. and Jan, N., 2020. Software Cost Estimation: Algorithmic and Non-Algorithmic Approaches. International Journal of Data Science and Advanced Analytics (ISSN 2563-4429), 2(2), pp.1-5.

[5] Kaushik, A., Chauhan, A., Mittal, D. and Gupta, S., 2012. COCOMO estimates using neural networks. International Journal of Intelligent Systems and Applications, 4(9), pp.22-28.

[6] Singh, B.K., Tiwari, S., Mishra, K.K. and Punhani, A., 2021. Extended COCOMO: robust and interpretable neuro-fuzzy modelling. International Journal of Computational Vision and Robotics, 11(1), pp.41-65.

[7] Coelho, E. and Basu, A., 2012. Effort estimation in agile software development using story points. International Journal of Applied Information Systems (IJAIS), 3(7).

[8] Bedi, R.P.S. and Singh, A., 2017. Software Cost Estimation using Fuzzy Logic. Indian Journal of Science and Technology, 10, p.3.

[9] Singh, B.K. and Misra, A.K., 2012. Software effort estimation by genetic algorithm tuned parameters of modified constructive cost model for nasa software projects. International Journal of Computer Applications, 59(9).

[10] Nassif, A.B., Azzeh, M., Capretz, L.F. and Ho, D., 2016. Neural network models for software development effort estimation: a comparative study. Neural Computing and Applications, 27(8), pp.2369-2381.

[11] Tayyab M.R., Usman M., Ahmad W. (2018) A Machine Learning Based Model for Software Cost Estimation. In: Bi Y., Kapoor S., Bhatia R. (eds) Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016. IntelliSys 2016. Lecture Notes in Networks and Systems, vol 16. Springer, Cham.

[12] Sakhrawi, Z., Sellami, A. & Bouassida, N. Software enhancement effort estimation using correlation-based feature selection and stacking ensemble method.

[13] Kumawat P., Sharma N. (2019) Design and Development of Cost Measurement Mechanism for Re-Engineering Project Using Function Point Analysis. In: Kamal R., Henshaw M., Nair P. (eds) International Conference on Advanced Computing Networking and Informatics. Advances in Intelligent Systems and Computing, vol 870. Springer, Singapore.

[14] J. A. Khan, S. U. R. Khan, T. A. Khan and I. U. R. Khan, "An Amplified COCOMO-II Based Cost Estimation Model in Global Software Development Context," in IEEE Access, vol. 9, pp. 88602-88620, 2021.

[15] P. Keil, D. J. Paulish, and R. S. Sangwan, "Cost estimation for global software development," in Proc. Int. Workshop Econ. Driven Softw. Eng. Res. (EDSER), 2006, pp. 7–10.

[16] Menzies T, Brady A, Keung J, Hihn J, Williams S, El-Rawas O, Green P, Boehm B. Learning project management decisions: a case study with case-based reasoning versus data farming. IEEE Transactions on Software Engineering. 2013 Sep 16;39(12):1698-713.

[17] D. Nandal and O. P. Sangwan, "Software cost estimation by optimizing COCOMO model using hybrid BATGSA algorithm," Int. J. Intell. Eng. Syst., vol. 11, no. 4, pp. 250–263, 2018.

[18] A. B. Nassif, M. Azzeh, A. Idri, and A. Abran, "Software development effort estimation using regression fuzzy models," Comput. Intell. Neurosci., vol. 2019, pp. 1–17, Feb. 2019.

[19] Zaidi SA, Katiyar V, Abbas SQ (2017) Development of a framework for software cost estimation: design phase. Int J Tech Res Appl 5(2):68–72.

[20] Reena, Bhatia PK (2017) Application of genetic algorithm in software engineering: a review. Int Refereed J Eng Sci 6(2):63–69.

[21] V. Venkataiah, R. Mohanty, M. Nagaratna, Prediction of software cost estimation using spiking neural networks, in: Smart Intell. Comput. Appl. Smart Innov. Syst. Technol., Springer, Singapore, 2019, pp. 101–112, http: //dx.doi.org/10.1007/978-981-13-1927-3_11.

[22] V. Venkataiah, R. Mohanty, M. Nagaratna, Prediction of software cost estimation using spiking neural networks, Smart Innov. Syst. Technol. 105 (2019) 101–112, http://dx.doi.org/10.1007/978-981-13-1927-3_11.

[23] S. Kumari, S. Pushkar, Cuckoo search based hybrid models for improving the accuracy of software effort estimation, Microsyst. Technol. 24 (2018) 4767–4774, http://dx.doi.org/10.1007/s00542-018-3871-9.

[24] M. Pandey, R. Litoriya, P. Pandey, Validation of existing software effort estimation techniques in context with mobile software applications, Wirel. Pers. Commun. 110 (2020) 1659–1677, http://dx.doi.org/10.1007/s11277-019-06805-0.

[25] S. Goyal, P.K. Bhatia, Feature selection technique for effective software effort estimation using multi-layer perceptrons, Lect. Notes Electr. Eng. 605 (2020) 183–194, http://dx.doi.org/10.1007/978-3-030-30577-2_15.

[26] A.J. Singh, M. Kumar, Comparative analysis on prediction of software effort estimation using machine learning techniques, SSRN Electron. J. (2020) 1–6, http://dx.doi.org/10.2139/ssrn.3565822.

[27] M. Qin, L. Shen, D. Zhang, L. Zhao, Deep learning model for function point based software cost estimation -an industry case study, in: Proc. - 2019 Int. Conf. Intell. Comput. Autom. Syst. ICICAS 2019, 2019, pp. 768–772, http://dx.doi.org/10.1109/ICICAS48597.2019.00165.

[28] V. Resmi, S. Vijayalakshmi, Kernel fuzzy clustering with output layer self-connection recurrent neural networks for software cost estimation, J. Circuits, Syst. Comput. 29 (2019) 1–17, http://dx.doi.org/10.1142/S0218126620500917.

[29] M. Choetkiertikul, H.K. Dam, T. Tran, T. Pham, A. Ghose, T. Menzies, A deep learning model for estimating story points, IEEE Transactions on Software Engineering, 45 (2019), 637–656, http://dx.doi.org/10.1109/TSE.2018.2792473.

[30] Dragicevic, S., Celar, S., & Turic, M. (2017). Bayesian network model for task effort estimation in agile software development. Journal of Systems and Software, 127, 109- 119. DOI: 10.1016/j.jss.2017.01.027.

[31] http://promise.site.uottawa.ca/SERepository/datasets/cocomonasa_v1.arff

[32] Stanley, K.O., Miikkulainen, R., 2002a. Evolving neural networks through augmenting topologies. Evolutionary Computation 10 (2), 99–127.

[33] X Chhabra, S., Singh, H. Optimizing design parameters of fuzzy model based COCOMO using genetic algorithms. Int. j. inf. tecnol. 12, 1259–1269 (2020). https://doi.org/10.1007/s41870-019-00325-7.

[34] A. A. Fadhil, R. G. H. Alsarraj and A. M. Altaie, "Software Cost Estimation Based on Dolphin Algorithm," in IEEE Access, vol. 8, pp. 75279-75287, 2020, doi: 10.1109/ACCESS.2020.2988867.

Incorporation of Computational Thinking Practices to Enhance Learning in a Programming Course

Leticia Laura-Ochoa, Norka Bedregal-Alpaca
Universidad Nacional de San Agustin de Arequipa, Arequipa, Peru

***Abstract*—The development of computational thinking skills is essential for information management, problem-solving, and understanding human behavior. Thus, the aim of the experience described here was to incorporate computational thinking practices to improve learning in a first Python programming course using programming tools such as PSeInt, CodingBat, and the turtle graphic library. A quasi-experimental methodological design was used in which the experimental and control groups are in different academic semesters. Exploratory mixed research was carried out. The control and experimental group consisted of 41 and 36 students, respectively. The results show that with the use of support programming tools, such as PSeInt, CodingBat, Python turtle graphic library, and the incorporation of computational thinking practices, the experimental group students obtained better learning results. It is concluded that student performance and motivation in university programming courses can be improved by using proper tools that help the understanding of programming concepts and the skills development related to computational thinking, such as abstraction and algorithmic thinking.**

***Keywords*—Programming tools; computational thinking; algorithmic thinking; motivation; abstraction**

I. Introduction

Computational Thinking (CT) is a fundamental skill for all students [1]. In [2], CT has been found to involve abstraction, algorithmic thinking, automation, decomposition, debugging, and generalization. In addition, the formation and development of algorithmic thinking in higher education students is a requirement of the information society, as it provides them with instruments to solve problems of everyday life [3] and get a solution through a series of steps [4]. It is a fundamental skill that students develop when they learn to program [5]. Also, computer programming involves other skills like logical reasoning and creativity in problem-solving.

However, learning computer programming for novice students is considered a challenge for educators, since a decrease in students' interest and motivation to learn programming courses has been noted [6]. The learning process can be complicated and demanding, difficult to master for novice programmers [6][7]. Computer programming courses are considered the most difficult courses in which undergraduate students do not usually succeed [8]; in [9] explain that the content of an introductory programming course emphasizes more on learning the syntax and semantics of the programming language. In addition, programming courses should not only focus on teaching students to write code but should also include the development of skills related to computational thinking, such as algorithmic thinking, logic, and problem-solving [4].

According to [10], programming courses introduce a programming language and the computer science thinking way. Furthermore, programming exposes students to computational thinking, because it requires problem-solving using computer science concepts such as abstraction and decomposition [11]. For [12][13], CT has begun to influence various disciplines and professions, in addition to all science and engineering disciplines, making it necessary to include it in general education.

On the other hand, high dropout rates are found in introductory programming courses [14], one of the main reasons being the lack of students' motivation [15]. Since computer programming requires constant effort and practice, it is important to keep students motivated [16], to get their predisposition to continue learning and improve their learning.

Consequently, the problem we found is that traditional methods used to teach programming courses to novice students, based on syntax and semantic content of the programming language can demotivate students to continue learning programming courses, causing low performance in their learning and even dropout.

In this context, the aim of the experience described here was to select the proper programming language, tools, and teaching strategies to teach introductory programming courses, so that students improve their learning outcomes, develop skills related to computational thinking, learn to program, and increase their motivation towards the subject of programming.

The rest of the paper is organized as follows: Section II provides some related works proposed in the literature. Section III describes the conceptual framework on algorithmic thinking, abstraction, and Python. Section IV explains the overview of the methodology. Section V presents a detailed description of the experience of incorporating computational thinking practices in the programming course. Section VI shows the results of applying programming tools and computational thinking practices to improve student performance. Section VII discusses the results obtained. Section VIII presents the conclusions and future work.

II. Related Work

In the work of [17], they present the use of the ADRI (Approach, Deployment, Result, Improvement) approach in the teaching and learning process of an introductory programming course, for which they redesigned their course materials and

developed an editor so that students can complete the required stages of the approach, managing to improve student learning outcomes compared to previous semesters, focusing on problem-solving strategies as well as programming knowledge. Additionally, ADRI's approach and editor reduced failure and dropout rates.

In [18], they present a teaching approach based on four components: The use of the Python programming language, project-oriented and problem-based learning methodologies, multimedia resources available on virtual platforms, and evaluation rubrics. The approach used improved the academic performance of the students, which is evidenced in the grades obtained, and the dropout rates were reduced. The results obtained suggest Python as a proper programming language for students of a first introductory programming course, due to its simplicity in syntax and code debugging, in addition to the use of other pedagogical strategies that support the learning process.

In the work of [3], they carry out an analysis of the scientific literature considering definitions, main properties, and characteristics of algorithmic thinking. They then present a universal sequence of algorithm development, involving different types of thinking such as abstract, conceptual, logical, constructive, and figurative. They carried out a survey in which the participants demonstrated a low level of understanding about algorithms, algorithmic thinking, and its usefulness in daily life and professional activity, so they end that algorithmic thinking is important for any higher education subject, not only in information and communication technologies (ICT) area and consider it as a new dimension of learning in higher education.

In [9], they introduce a new teaching approach focusing on algorithmic thinking skills besides the knowledge of the syntax and semantics of a programming language in an introductory programming course, using techniques of flowchart and pseudocode. Their results show that the ADRI approach promotes the three-step approach (Problem statement → Solution plans → Code) to solve a problem, fosters programming knowledge, as well as problem-solving strategies, promoting algorithmic thinking.

In the work of [19], they describe the design and implementation of an introductory computational thinking course to teach programming to high school students with activities that take place in a web-based programming environment that uses a variant of the Haskell language, promoting higher-order thinking. They address the need for computational thinking courses geared toward all students, not just future software developers, by making connections between learning programming with science and math. Most of the students who participated in the course considered it difficult; but there was an overall positive reception from the students, who learned the language and the general principles of programming, logic, and modeling. They find that courses like Python typically do not focus on computational thinking and follow traditional syntax-oriented approaches to teaching programming, with little connection to science and math.

III. Conceptual Framework

A. Algorithmic Thinking

Algorithmic thinking is essential in comprehensive general education and programming is a way to teach the basic principles of algorithmic thinking from the beginning [10], it is important in higher education, to develop algorithms in the context of the future profession and everyday life in the modern information society [3]. Also, it is considered a significant component of the cognitive competencies of the future engineer because the algorithmic activity allows forming adequate algorithmic skills, through which students develop techniques of mental actions such as generalization, classification, analogy, the establishment of patterns and logical reasoning, which are the main components of algorithmic thinking [20]. Therefore, it is advisable to promote the development of algorithmic thinking skills through programming in the different disciplines and professions, besides careers related to computing.

According to [5][4][21], algorithmic thinking consists of a clear definition of the steps to reach a solution, thinking in terms of instruction sequences and rules that lead to problem-solving or understanding of situations. It is an important aspect of computational thinking [22], its main properties include discretion, abstraction, formality, integrity, and effectiveness [3].

B. Abstraction

Abstraction, efficiency, and algorithms are considered vital "mental tools" for computational thinking [23]. According to [13], abstraction is the most important and high-level thought process in computational thinking.

Abstraction is a key skill for computing, fundamental for mathematics and engineering in general [24], it involves reducing unnecessary details, eliminating complexity, choosing the correct detail to hide, and thus the problem is easier and understandable without missing anything important [21][5]. Therefore, it allows developing a potential solution by eliminating details of the problem [23]. For [25], abstract thinking is the ability to abstract the properties of objects that are relevant to a study.

Furthermore, abstraction allows defining patterns, generalizing by capturing common essential properties from instances, and parameterization [13]. Without abstraction, students tend to get overwhelmed with details and feel frustrated with the programming process [23], so the development of this skill is necessary, applicable in programming, mathematics, and the different disciplines.

C. Python

The main professional programming languages are based on text such as C, Python, Java [26]. Among these, the use of Python makes it easier for novice students to engage in the main features of computational thinking, mainly due to its basic syntax, dynamic typing (declaring variables is not required), structured and indented writing [4]. Its use is suitable because it includes turtle graphics libraries that allow a smooth transition from Logo to Python [10], which allows focusing on

concepts without a long introduction to the syntactic details of the language [27].

In addition, Python is a high-level programming language, easy to learn, free, and with documentation available on the Web [4].

IV. Methodology

The methodological design used was quasi-experimental with experimental and control groups located in different semesters, so the selection of its members was not random. Exploratory mixed research was carried out.

In the experience, the experimental group consisted of 36 students enrolled in the Programming course, group A, of the 2019-A academic period of the Professional School of Mechanical Engineering of the Universidad Nacional de San Agustin de Arequipa (Peru). In this group, there were 34 male students (94%) and 2 female students (6%). The control group was the students of group A who completed the Programming course in the 2018-A academic period, made up of 41 male students (100%).

The Programming course at the Professional School of Mechanical Engineering of the Universidad Nacional de San Agustin de Arequipa - Peru, is given in the third academic semester and it is developed for 17 weeks. It has 3 hours a week (1 theoretical hour and 2 laboratory hours), it is equivalent to 2 credits and the Python programming language is used.

In the control group, tools such as DFD were used to create data flow diagrams and PSeInt for pseudocode, before the use of the Python programming language; CodingBat and turtle graphic library were not used. A greater preference was also observed for the use of the PSeInt tool concerning the DFD tool, so in the experimental group only PSeInt was used and the use of the CodingBat tool and the Python turtle graphic library were incorporated with an approach oriented to computational thinking practices, in addition to the Python programming language.

Data collection was done from the students' grades obtained in their evaluations of the programming course, before (control group) and after the experiment (experimental group). Direct observations were also made during the programming activities.

To measure success, a comparison of the grades obtained by the students of the control and experimental groups was made, to check if there is an improvement in the students' performance of the experimental group. The results were validated by statistical analysis using SPSS Statistic V25 software, to determine if there is a statistically significant improvement.

V. Description of the Experience

This work describes our experience in the use of PSeInt, CodingBat, and Python turtle graphical library to motivate and reinforce students' learning of programming concepts and develop skills related to computational thinking in a first programming course with Python.

In the programming course, students learned topics such as sequential statements, conditionals, loops, functions, structured types, object-oriented programming.

In the 2019-A academic period, students began learning to create algorithms to solve problems using the PSeInt tool (Fig. 1), with which they developed algorithmic thinking skills, logic, and problem-solving strategies using pseudocode.

With PSeInt, the students were able to execute algorithms in an automated way to test their solution proposals and verify results, analyzing the errors in the logic, which allowed the student to practice automation and debugging.

Then, the Python programming language was taught. Initially, they were asked to perform the same exercises developed in PSeInt, to pass their created algorithms to a computer program using the Python programming language.

Students learned the syntax and semantics of the Python programming language, practiced coding, running programs, checking results, parsing, and fixing syntax errors. Automation and debugging were also present.

To improve students' programming skills, the online code practice tool called CodingBat [28] was used, which presents some examples with solutions available for students to practice coding and executing programs in Python, allowing them to check your answer or see other ways to solve the same problem, plus there are several exercises to solve with hints available, using conditionals, loops, strings, and lists. Fig. 2 shows some exercises that were solved by the students using Boolean logic and conditionals in CodingBat, which provided them with more opportunities to practice their programming constructions, as well as reinforce computational concepts such as sequential and conditional instructions.

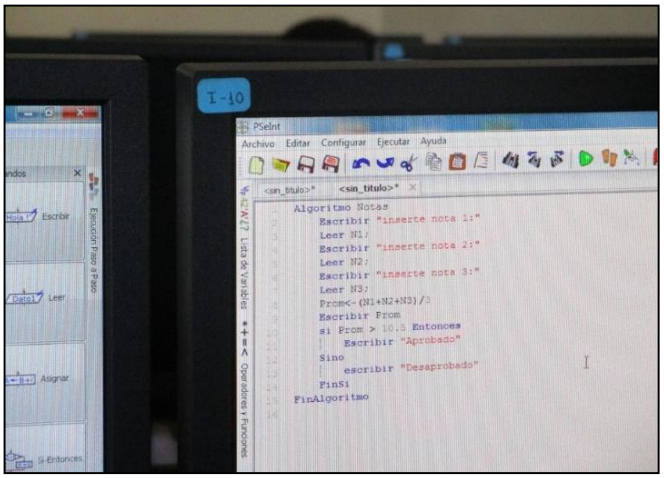


Fig. 1. Creation of Algorithms through Pseudocode using the PSeInt Tool.

Fig. 2. Carrying out Python Logic-1 Exercises in CodingBat.

Students learned to create graphics using Python turtle graphical library, they began by drawing geometric figures using sequential instructions and functions, then drawing different shapes and patterns (Fig. 3, 4, 5) incorporating repetitive instructions.

Fig. 3 shows geometric exercises performed by the students, where they apply iterations and functions from squares and rectangles to create different graphics with repetitive patterns.

Fig. 4 shows additional geometric exercises created from parallelograms and circles, where students apply their creativity and logic with the help of iterations and functions.

Fig. 5 shows an example, in which the problem is first decomposed using the functions parallelograms (to draw small rhombus by tracing lines of 65 pixels) and parallelogram (to draw large rhombus by tracing lines of 100 pixels). Through abstraction and generalization, repetitive patterns were identified to create the graphics, and then their abstractions were improved by proposing new solution strategies using parameters, which allowed reducing the two functions that drew the rhombuses of different sizes into a single function called parallelograms(n). The function parallelograms(n) groups instruction patterns to draw shapes by drawing lines according to the number of pixels specified in the parameter n, this function is reused from the main function called main.

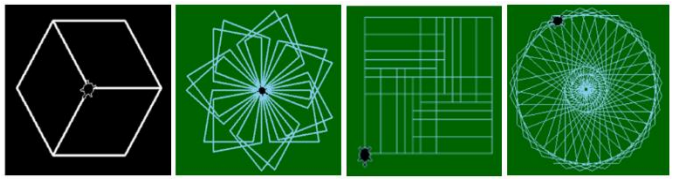


Fig. 3. Drawing Geometric figures such as Squares, Rectangles with different Angles of Rotation using Iterations and Functions.

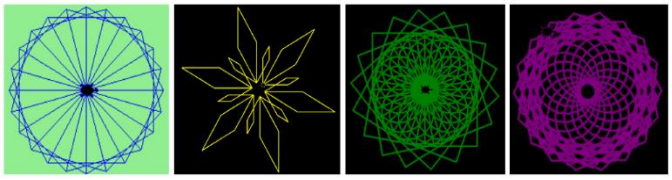


Fig. 4. Drawing Geometric figures Like Parallelograms, Circles using Iterations and Functions.

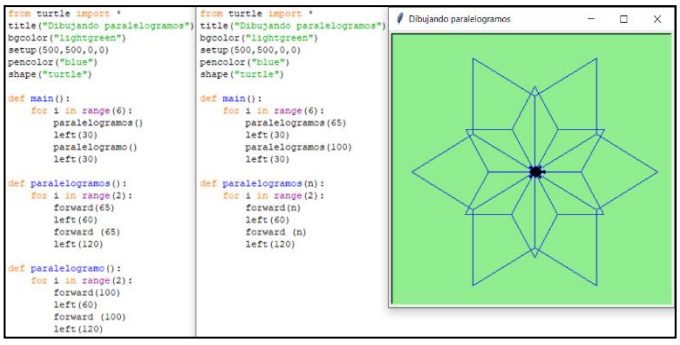


Fig. 5. Using Functions, Loops, Parallelograms of different Sizes (65 and 100 Pixels) with Rotation Angles of 60 and 120 Degrees.

In addition to reinforcing programming concepts such as sequential instructions, loops, and functions, students gained computational thinking practices such as decomposition, iteration, and abstraction that enabled them to recognize repeating patterns.

VI. Results

In the academic period 2019-A, students obtained an average grade of 16.75 in their first exam. Then, in the evaluation with Python, they obtained an average grade of 13.86 in their second exam. Finally, CodingBat and the turtle graphic library were used to reinforce and motivate them in their learning process, obtaining an average grade of 14.97, which improved their grade using only PSeInt for algorithms creation and the Python programming language. Table I shows the average of the grades obtained in the first, second and third exams of group A of the Programming course taught in the academic periods 2018-A and 2019-A, which range from 0 to 20.

Fig. 6 shows the average grades evolution for Exam 1, 2, and 3 in the academic periods 2018-A and 2019-A, showing an improvement for Exam 1 and Exam 3 in 2019-A.

The use of PSeInt, CodingBat, and turtle graphic library has shown an improvement in the students' grades in their average grades. Table II shows the global grade average of group A in 2018-A and 2019-A, which range from 0 to 20.

Fig. 7 shows the global grade average for the Programming course in the academic periods 2018-A and 2019-A, showing an improvement in 2019-A.

TABLE I. Average Grades for Exam 1, 2 and 3

Academic period	*Exam 1*	*Exam 2*	*Exam 3*
2018-A	14.34	13.66	13.37
2019-A	16.75	13.86	14.97

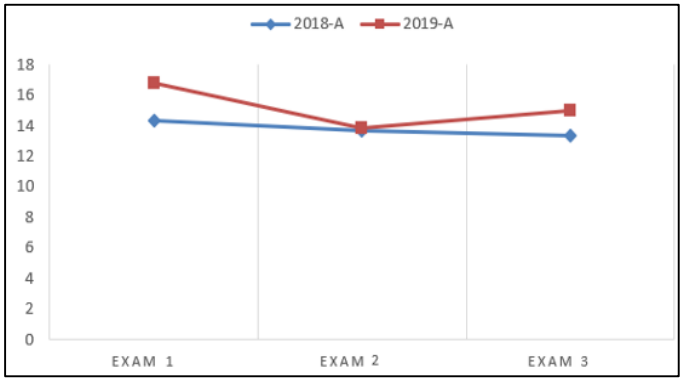


Fig. 6. Average Grade Result for Exam 1, 2, and 3 of the Programming Course by Academic Year.

TABLE II. Global Grade Average

Academic period	*Global average*
2018-A	14.3
2019-A	15.08

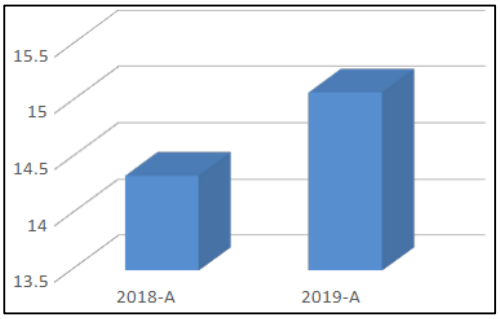


Fig. 7. Global Grade Average for the Programming Course by Academic Year.

In our experience, we have observed that novice students are motivated by using Python turtle graphic library to create their drawings and improve their abstractions, as well as develop skills related to computational thinking. CodingBat allowed them to improve their programming skills by practicing their coding in Python, thereby improving their final grades in the programming course.

SPSS Statistic V25 software was used for the statistical analysis of the results. Table III has some descriptive measures for the grades obtained by students in the years 2018 and 2019. The average of the grades for the year 2018 is 14.3 and with a 95% confidence level, it can be stated that range between 13.86 and 14.82, while the average of the grades for the year 2019 is 15.08 and with the same 95% confidence level it can be stated that range between 14.31 and 15.85. Consequently, it could be assumed that the grades obtained in 2019 were better than those of 2018.

To determine if the difference was statistically significant, as a requirement, it was necessary to verify that the conditions of normality and heteroskedasticity were met. With a significance level of 5% (0.05), the Kolmogorov-Smirnov test was applied, obtaining the results of Table IV, as the p-value (Sig.) is greater than the significance, and the grades distribution normality is accepted.

When applying the t-test for independent samples, the results shown in Table V were obtained.

In Levene's test, as the p-value of 0.27 is greater than the significance, then it was possible to affirm that the assumptions of normality and heteroskedasticity were met, therefore, it was possible to apply the t-test for the means difference. Table V shows that the bilateral p-value is 0.093, so the unilateral value is 0.046, which is less than the significance; consequently, the means equality hypothesis is rejected.

TABLE III. DESCRIPTIVE MEASURES FOR THE GRADES OBTAINED IN THE YEARS 2018 AND 2019

Grade	2018	2019
Mean	14.3415	15.0833
Standard deviation	1.52659	2.27251
Standard error of the mean	0.23841	0.37875
N	41	36
CI 95% lower limit	13.86	14.31
CI 95% upper limit	14.82	15.85

TABLE IV. RESULTS OF THE KOLMOGOROV-SMIRNOV NORMALITY TEST

Year	Kolmogorov-Smirnov[a]		
	Statistic	*df*	*Sig.*
2018	,149	41	,052
2019	,126	36	,163

Lilliefors Significance Correction

TABLE V. INDEPENDENT SAMPLES T-TEST

	Levene's test for equality of variances		t-test for equality of means	
	F	*Sig.*	*t*	*Sig. (2-tailed)*
Equal variances assumed	5.120	0.27	-1,699	,093
Equal variances not assumed			-1,658	,103

VII. DISCUSSION

Statistically, it is possible to state that the difference found between the grades obtained in 2018 and 2019 were statistically different and, with a confidence level of 95% it can be stated that the grades obtained by the students in 2019 were better than those obtained in 2018.

The use of support tools such as PSeInt, CodingBat, and Python turtle graphic library have increased students' motivation and performance in a first programming course with Python, because an approach oriented to computational thinking practices was also followed. As indicated by [29], computer programming is the main demonstration of computational thinking skills. However, they tend to follow syntax-oriented programming teaching approaches, without focusing on computational thinking, with few connections to mathematics and science [19]. Furthermore, in the work of [18], they indicate that the Python programming language is suitable for introductory programming courses because of its simple syntax and ease for code debugging, they also point out that it is necessary to consider other aspects such as pedagogical strategies that allow improving the programming teaching-learning process.

According to [5], algorithmic thinking is a fundamental skill that students acquire when they learn to program, developing the ability to think in terms of sequences and rules to solve problems. In the work of [9], they used flowcharts and pseudocode for novice students to propose solutions to problem statements, which promote algorithmic thinking skills. Similarly, in our work, we use the PSeInt tool for students to develop algorithmic thinking skills, logic, and problem-solving strategies by creating algorithms with pseudocode. In addition, because it is a tool that allows algorithms execution in an automated way, the students were able to test their solution proposals, find and resolve errors in logic, practicing automation and debugging, which are part of the computational thinking skills found in the work of [2].

We consider that the use of support tools such as CodingBat is essential so that students can practice their programming constructs and improve their problem-solving skills in the programming language used in the course.

Likewise, in the work of [17], they express the importance of practicing programming skills by dedicating more time and focusing on problem-solving strategies.

According to the experience described, the Python turtle graphic library allowed the acquisition of computational practices such as abstraction, decomposition, iteration, and debugging, which correspond to the computational thinking practices defined by [30] and adapted in the work of [31]. In addition, in the programming course, students developed skills such as algorithmic thinking, automation, and generalization, which are part of the computational thinking skills identified in five articles highlighted in the work of [2], which are abstraction, algorithmic thinking, automation, decomposition, debugging, and generalization. On the other hand, in the programming course, concepts such as sequential instructions, loops, conditionals were learned, which correspond to the concepts of computational thinking considered in the work of [31].

We agree with [32], in the sense that programming environments with graphic components allow the acquisition of computational thinking practices through programs creation, which is attractive to them, helping students develop a positive attitude towards programming.

Currently, technology-mediated training processes are becoming increasingly flexible and collaborative; therefore, problem-solving activities can be involved [33] that involve cooperative learning techniques [34], such as the programming of a robotic hand. Consequently, the student would not only be developing computational thinking but also critical spirit, creativity, and collaborative work.

VIII. Conclusion

In this study, we considered two semesters with different students. Semester 2018-A with 41 students from group A, served as a control group, where we used the DFD and PSeInt tools earlier to teach programming using Python, while in semester 2019-A, with 36 students from group A, we started with the PSeInt tool before teaching Python, then the students' learning of Python programming was reinforced with the CodingBat tool and turtle graphic library. We examined the grades obtained in the midterm exams and the global average of the programming course, where an improvement in the grades in the second experimental group with the support tools used in the course is evidenced, indicating that they acquired better programming skills and therefore better performance. In addition, we observed that students are motivated by using the Python turtle graphic library that reinforces their learning of sequential instructions, loops, functions, and allows the development of skills related to computational thinking such as algorithmic thinking, decomposition, iteration, and abstraction. The experience of this work can serve as a reference for educators interested in approaches oriented to computational thinking practices in programming teaching.

As future work, we consider investigating computational thinking measurement evaluations to be applied after following a programming teaching approach oriented to computational thinking practices.

Acknowledgment

The authors' thanks are expressed to the National University of San Agustin de Arequipa for the support received in the realization of the proposal and the results are expected to benefit the institution.

References

[1] Q. Li, “Computational thinking and teacher education: An expert interview study,” Human Behavior and Emerging Technologies, vol. 3, no. 2, pp. 324-338, 2021.

[2] S. Bocconi, A. Chioccariello, G. Dettori, A. Ferrari and K. Engelhardt, “Developing computational thinking in compulsory education - Implications for policy and practice,” in JRC Science for Policy Report, 2016.

[3] M. F. Byrka, A. V. Sushchenko, A. V. Svatiev, V. M. Mazin and O. I. Veritov, “A New Dimension of Learning in Higher Education: Algorithmic Thinking,” Propósitos y Representaciones, vol. 9, no. SPE2, pp. 990, 2021.

[4] F. Buitrago Flórez, R. Casallas, M. Hernández, A. Reyes, S. Restrepo and G. Danies, “Changing a generation’s way of thinking: Teaching computational thinking through programming,” Review of Educational Research, vol. 87, no. 4, pp. 834-860, 2017.

[5] A. Csizmadia, P. Curzon, M. Dorling, S. Humphreys, T. Ng, C. Selby and J. Woollard, “Computational thinking-A guide for teachers,” Computing at School, 2015.

[6] M. Piteira and C. Costa, “Computer programming and novice programmers,” in Proceedings of the Workshop on Information Systems and Design of Communication pp. 51-53, 2012.

[7] M. Karaliopoulou, I. Apostolakis and E. Kanidis, “Perceptions of Informatics Teachers Regarding the Use of Block and Text Programming Environments,” European Journal of Engineering Research and Science, pp. 11-18, 2018.

[8] B. Özmen and A. Altun, “Undergraduate students' experiences in programming: difficulties and obstacles,” Turkish Online Journal of Qualitative Inquiry, vol. 5, no. 3, pp. 1-27, 2014.

[9] S. I. Malik, M. Shakir, A. Eldow and M. W. Ashfaque, “Promoting Algorithmic Thinking in an Introductory Programming Course,” International Journal of Emerging Technologies in Learning, vol. 14, no. 1, 2019.

[10] J. Hromkovic, T. Kohn, D. Komm and G. Serafini, “Algorithmic thinking from the start,” Bulletin of EATCS, vol. 1, no. 121, 2017.

[11] S. Y. Lye and J. H. L. Koh, “Review on teaching and learning of computational thinking through programming: What is next for K-12?,” Computers in Human Behavior, vol. 41, pp. 51-61, 2014.

[12] J. M. Wing, “Computational thinking,” Communications of the ACM, vol. 49, no. 3, pp. 33-35, 2006.

[13] J. M. Wing, “Computational thinking: What and why. The Link,” News from the School of Computer Science at Carnegie Mellon University, 2011.

[14] C. Chen, P. Haduong, K. Brennan, G. Sonnert and P. Sadler, “The effects of first programming language on college students’ computing attitude and achievement: a comparison of graphical and textual languages,” Computer Science Education, vol. 29, no. 1, pp. 23-48, 2019.

[15] P. Kinnunen and L. Malmi, “Why students drop out CS1 course?,” in Proceedings of the second international workshop on Computing education research, pp. 97-108, 2006.

[16] A. Settle, A. Vihavainen and J. Sorva, “Three views on motivation and programming,” in Proceedings of the 2014 conference on Innovation & technology in computer science education, pp. 321-322, 2014.

[17] S. Iqbal Malik and J. Coldwell-Neilson, “Impact of a new teaching and learning approach in an introductory programming course,” Journal of Educational Computing Research, vol. 55, no. 6, pp. 789-819, 2017.

[18] O. Solarte Pabón and L. E. Machuca Villegas, “Fostering Motivation and Improving Student Performance in an Introductory Programming Course: An Integrated Teaching Approach,” Revista EIA, vol. 16, no. 31, pp. 65-76, 2019.

[19] F. Alegre, J. Underwoood, J. Moreno and M. Alegre, "Introduction to Computational Thinking: a new high school curriculum using CodeWorld," in Proceedings of the 51st ACM Technical Symposium on Computer Science Education, pp. 992-998, 2020.

[20] M. Kovalchuk, A. Voievoda and E. Prozor, "Algorithmic Thinking as the Meaningful Component of Cognitive Competencies of the Future Engineer," Universal Journal of Educational Research, vol. 8 (11B), pp. 6248-6255, 2020.

[21] P. Curzon, M. Dorling, T. Ng, C. Selby and J. Woollard, "Developing computational thinking in the classroom: a framework," Computing at School, 2014.

[22] M. Romero, A. Lepage and B. Lille, "Computational thinking development through creative programming in higher education," International Journal of Educational Technology in Higher Education, vol. 14, no. 1, pp. 1-15, 2017.

[23] J. A. Qualls, M. M. Grant and L. B. Sherrell, "CS1 students' understanding of computational thinking concepts," Journal of Computing Sciences in Colleges, vol. 26, no. 5, pp. 62-71, 2011.

[24] J. Kramer, "Is abstraction the key to computing?," Communications of the ACM, vol. 50, no. 4, pp. 36-42, 2007.

[25] M. Zapata-Ros, "Pensamiento computacional: Una nueva alfabetización digital," Revista de Educación a Distancia (RED), vol. 46, no. 4, 2015.

[26] M. Kölling, N. C. Brown and A. Altadmri, "Frame-based editing: Easing the transition from blocks to text-based programming," in Proceedings of the Workshop in Primary and Secondary Computing Education, pp. 29-38, 2015.

[27] J. Hromkovič, T. Kohn, D. Komm and G. Serafini, "Combining the power of python with the simplicity of logo for a sustainable computer science education," in International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, Springer, Cham, pp. 155-166, 2016.

[28] N. Parlante. (2017). CodingBat code practice [Online]. Available: https://codingbat.com/python

[29] M. Román-González, J. C. Pérez-González, J. Moreno-León and G. Robles, "Can computational talent be detected? Predictive validity of the Computational Thinking Test," International Journal of Child-Computer Interaction, vol. 18, pp. 47-58, 2018.

[30] K. Brennan and M. Resnick, "New frameworks for studying and assessing the development of computational thinking," in Proceedings of the 2012 annual meeting of the American Educational Research Association, 2012.

[31] F. Luo, P. D. Antonenko and E. C. Davis, "Exploring the evolution of two girls' conceptions and practices in computational thinking in science," Computers & Education, vol. 146, pp. 103759, 2020.

[32] L. Laura-Ochoa and N. Bedregal-Alpaca, "Análisis de entornos de programación para el desarrollo de habilidades del pensamiento computacional y enseñanza de programación a principiantes," Revista Ibérica de Sistemas e Tecnologias de Informaçao, no. E43, pp. 533-548, 2021.

[33] V. Cornejo-Aparicio, S. Flores-Silva, N. Bedregal-Alpaca and D. Tupacyupanqui-Jaén, "Capstone courses under the PBL methodology approach, for engineering," 2019 IEEE World Conference on Engineering Education (EDUNINE), 2019, DOI: 10.1109/EDUNINE.2019.8875803.

[34] N. Bedregal-Alpaca, V. Cornejo-Aparicio, A. Padron-Alvarez and E. Castañeda-Huaman, "Design of cooperative activities in teaching-learning university subjects: Elaboration of a proposal," International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, 2020, DOI: 10.14569/IJACSA.2020.0110445.

Detecting and Fact-checking Misinformation using "Veracity Scanning Model"

Yashoda Barve[1], Jatinderkumar R. Saini[2]*, Ketan Kotecha[3], Hema Gaikwad[4]
Suryadatta College of Management, Information Research & Technology, Pune, India[1]
Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India[3]
Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India[2, 4]

***Abstract*—The expeditious flow of information over the web and its ease of convenience has increased the fear of the rampant spread of misinformation. This poses a health threat and an unprecedented issue to the world impacting people's life. To cater to this problem, there is a need to detect misinformation. Recent techniques in this area focus on static models based on feature extraction and classification. However, data may change at different time intervals and the veracity of data needs to be checked as it gets updated. There is a lack of models in the literature that can handle incremental data, check the veracity of data and detect misinformation. To fill this gap, authors have proposed a novel Veracity Scanning Model (VSM) to detect misinformation in the healthcare domain by iteratively fact-checking the contents evolving over the period of time. In this approach, the healthcare web URLs are classified as legitimate or non-legitimate using sentiment analysis as a feature, document similarity measures to perform fact-checking of URLs, and incremental learning to handle the arrival of incremental data. The experimental results show that the Jaccard Distance measure has outperformed other techniques with an accuracy of 79.2% with Random Forest classifier while the Cosine similarity measure showed less accuracy of 60.4% with the Support Vector Machine classifier. Also, when implemented as an algorithm Euclidean distance showed an accuracy of 97.14% and 98.33% respectively for train and test data**.

***Keywords*—Document similarity; fact-checking; healthcare; incremental learning; misinformation; sentiment analysis**

I. Introduction

The exponential growth of the internet and World Wide Web (WWW) and its ease of convenience, has led to an information flow expeditiously. Social media, especially Facebook and Twitter have become major sources for information sharing. The expediency, diversified knowledge, and reasonable cost attract the users of the internet to access and share information online, leading to a rapid generation of information [1]. In the healthcare domain, an enormous volume of health and medical-related material is accessible online. It was observed that physicians choose the web as a valuable information resource for medical practice, education, or learning as well as decision support while patients surf the internet for information on diseases, infections, and their indications. For example, 65% of users prefer the internet to search health-related topics [2, 3, 4]. According to the survey in 2017, by Pew Research Center, 88% of American people have quick access to the internet at home and 81% of them get updates of news from the internet [5]. Therefore, it can be determined that the users make maximum usage of the internet for information access.

However, the material made available online doesn't guarantee quality as well as correctness. The credibility and veracity of information is a major concern as it may lead to the rampant spread of misinformation [3, 4]. Misinformation is inaccurate or incorrect information that can be verified with available facts. The misinformation or false information may appear in various forms like fake news, rumor, satire news, hoaxes, misinformation, disinformation, etc. This massive spread of misinformation over the web has detrimental effects on people's life [6].

Apart from the existing health crisis, the spread of the ubiquitous problem of misinformation poses additional health threats and presents another unprecedented issue to the world [7, 8]. This creates a severe effect on people's life and medical experts as well [9]. For example, during the recent Covid-19 pandemic, misinformation about ingesting fish tank cleaning products can cure the virus or 5G networks generate radiations that triggers the virus or statement like "coronavirus is just like the flu" or "coronavirus is an engineered bioweapon" had an impact on people that they started believing the misinformation. Such misinformation causes panic amongst citizens and may lead to death [5, 10]. During 2014, Ebola outbreak, misinformation on the web and social media about some products which can cure Ebola had led to deaths [2]. Another example, misconception about the measles, mumps, and rubella (MMR) vaccine producing autism had a negative societal impact. Therefore, detecting misinformation has become a necessity to provide timely, verified, and credible information to the users in a way that can benefit society as a whole. Failure to meet this requirement can promote the misuse of misinformation which has adverse effects [1, 5, 10].

Researchers have been passionate about finding solutions to the misinformation detection problem. For example, recently, big social media companies like Facebook, Twitter, and Google have developed machine learning and deep learning-based models to detect misinformation on Covid-19 related posts and ads. In this, Facebook reported that they have had identified and deleted around 50 million posts on Covid-19 while Google and Twitter have taken corrective actions to remove scammer ads on face masks, hand sanitizers, etc. [10]. However, simply detecting misinformation cannot guarantee the veracity or credibility of information. Hence, fact-checking has an increasing demand for veracity scanning of information that can classify information as true or false [11].

*Corresponding Author.

Fact-checking is assessing the truthiness of information that is under investigation in an attempt to identify whether the information is factual [11, 12]. Automatic fact-checking refers to checking the truthiness of the information repeatedly based on all available data and classifying it into True, False, Mostly True, Mostly False, and Half True. According to [12] the process of fact-checking involves identifying the context of the claim, identifying new and previously fact-checked claims, and performing fact-checking with existing verified. According to the literature, there are three main techniques to perform fact-checking based on the evidence used to determine the veracity of the information. The first is the reference approach, these are based on valid or recognized sources and claims which are fact-checked beforehand. Second, knowledge graph approaches, which are based on subject-predicate-object triples for fact-checking [2]. The third category is contextual approaches, which involve perceptive about societal and other context-related claims. Many researchers have developed models and fully automated tools like VERA and Claimbuster to fact-check claims in all three categories. The knowledge graphs and contextual approaches showed higher accuracy values on author-generated datasets but accuracy decreased on different datasets like FEVER and HeroX reaching between 50% and 65% whereas reference approaches succeeded between 77% and 82%. These results reveal that automated fact-checking is a challenging task to resolve fully [12].

The main challenges in the automated fact-checking process involve 1) unavailability of standard annotated datasets in a specific domain. For example, fact-checking websites like politifact.com or fact-check.org mainly focus on specific domains viz. politics, news, etc. Thus, obtaining comprehensive datasets from these websites in a certain domain is not possible. 2) Expert and human annotation is extremely time-consuming and costly. It was studied that the most reliable approach in misinformation detection is to perform human expert-based fact-checking of data. However, with the large volume of data and the haste with which the misinformation is generated and disseminated uncontrollably, manual fact-checking is become time-consuming and might not be able to stop the impact of misinformation in its early stages [7]. 3) Verifying the truthiness of contents with the knowledge base. Therefore, there is a pressing need to design a dynamic and automatic fact-checking model to detect and verify healthcare misinformation [7, 12, 13].

Hence, to deal with data drift occurring in the model and to detect misinformation by iteratively performing fact-checking the authors have proposed a Veracity Scanning Model (VSM) using a combination of techniques viz. incremental learning, sentiment analysis, and standard document similarity measures.

A hybrid approach of incremental learning, sentiment analysis, and document similarity can help to detect misinformation as well as perform fact-checking with already verified data, and also handle the newly arriving chunk of data on the web automatically. Following are the research objectives.

1) To develop a methodology to perform automatic fact-checking using standard document similarity measures viz. Euclidean Distance, Jaccard Distance, and Cosine Similarity and classify healthcare URLs as Legitimate or Non-Legitimate using Veracity Scanning Model (VSM).

2) To evaluate and validate the performance of the proposed model.

The remaining section of the paper is structured as follows: Section II discusses the literature, Section III explains the methodology and Section IV highlights the results and discussion followed by Section V conclusion and future enhancements.

II. Literature Survey

This section discusses various techniques used in the literature to tackle above mentioned challenges. Section A describes the reason behind using incremental learning, section B elaborates on misinformation detection techniques and section C focuses on fact-checking methods.

A. Incremental Learning Approach

The classic problem of false information or misinformation detection and fact-checking deals with the static data and does not consider the streaming nature of the data. The profile of information classified as true and false may change over time. This results in a phenomenon called concept drift of data drift. In the literature, the researchers have fingered such problems using techniques like ensemble learning, or incremental learning [14, 15]. The technique of ensemble learning involves dividing the data stream into small chunks and then training each of the data chunks with different classifiers and ultimately choosing the best classifier. These types of algorithms are recommended to handle sudden or rigorous concept drift and are not much suitable for incremental drift of data [14]. In [15] the authors have used ensemble learning technique with online Bagging with classifiers viz. multi-layer perceptron, Gaussian Naïve Bayes Hoeffding Tree. Incremental Learning (IL) techniques iteratively learn knowledge from newly arriving data without forgetting previously learned knowledge without retraining the model on a complete dataset. Thus, the necessity of the availability of whole labeled data vanishes. Hence, the incremental learning approach is considered to be more suitable to handle smooth concept drifts, and have better performance on efficiency [16, 17, 18 19]. In the literature, researchers have used incremental learning techniques to detect fake news using Artificial Neural Networks (ANN). However, ANNs suffer from catastrophic forgetting which lowers the performance of the model as data streams arrive [20]. The deep learning and neural network-based techniques can classify short text appearing sequentially but require large memory space and training time, thus reducing the performance of the model [19]. Therefore, a novel incremental approach of VSM can be efficiently used to classify the textual data of false information or misinformation.

B. Detecting Misinformation using Sentiment Analysis

Detecting misinformation has gained researchers' attention and is widely focused on politics and mass communication areas. However, less attention is paid to the healthcare domain. The healthcare-related misinformation is studied in five different categories mainly communicable diseases,

infections like Zika Virus, Ebola, influenza, etc., chronic non-communicable diseases, diet and nutrition, smoking, and water safety. The selection of the right features plays a key role in detecting misinformation. In the literature, researchers have focused on several types of features like syntactical, user-specific, image-specific, sentimental, etc. However, sentimental features are found to be the most effective in determining the percentage of misinformation in a document [6, 21]. In [22, 23] authors have focused on sentimental features for healthcare misinformation detection. Thus, in this research authors have considered the sentimental features as a central feature.

C. Fact-Checking using Incremental Learning

The baseline approach for automatic fact-checking using referencing is finding the resemblance among new statements with already fact-checked statements such as Jaccard Distance, Cosine Similarity, Euclidian Distance, Manhattan Distance, etc. [24]. However, the static model can't cope up with incremental data popping up over a period of time. Thus techniques like incremental learning should be adopted. Incremental learning is the process of adapting to the newly arriving data, without the need to reprocess the old instance but remembering previously learned knowledge. In a research incremental learning was adopted to identify new features and new classes as the documents evolve over the time period with the help of incremental neural network based on neural perceptron [25]. To classify documents based on security, a methodology consisting of the combination of incremental learning and similarity features was proposed. Incremental learning is achieved through documental representation and similarity is measured by fetching sentence features. The classification process is based on security labels of already classified documents [26]. In another research, incremental learning for Hierarchical Dirichlet Process (HDP) was used with partial supervision i.e. the training data contains a mixture of labeled and unlabeled documents. Incremental learning is considered for newly arriving data without referring to previously learned data and also maintaining the robustness and consistency of the model. It was observed that the partially labeled dataset makes an important contribution to achieving good accuracy. The model also introduces granular computing to handle unlabeled data [27]. Thus, to update the model automatically after the arrival of new data and consequently classify/cluster the newly arriving documents either into a fixed number of classes or identify new classes or generate new features for document classification or clustering, it is a good approach to combine incremental learning and document similarity techniques [24, 28, 29]. In the author's previous work [1], a fact-checking model was proposed to find misinformation in the healthcare domain. In this research, authors have proposed a new technique of threshold computation function to classify URLs as legitimate or non-legitimate along with incremental learning to deal with data drifts occurring over the period of time and perform fact-checking.

D. Potential Research Gaps Identified

Following is the summarized list of potential research gaps identified through extensive literature from section A, B and C:

1) The research work conducted previously does not tackle the problem of incremental data appearing at different interval of time while dealing with the misinformation detection problem.

2) To the best of the author's knowledge, detecting misinformation via fact-checking is not studied extensively in the literature.

3) Recent techniques in this area focus on extracting features from the text and classifying the text as true or false. However, the authors found that the veracity of information plays a significant role in the classification of misinformation.

III. Methodology

The Veracity Scanning Model (VSM) detects and performs fact-checking of healthcare data using an incremental learning approach. VSM consists of three main phases viz. Monitoring, Spotting, and Checking. These phases are generated based on the fact-checking process model defined in the literature [12]. The Monitoring phase consists of fetching healthcare URLs and generating sentimental Bag-of-Words (s-BoW). The spotting phase includes extracting features and detecting misinformation based on features extracted. The checking Phase consists of performing fact-checking and ultimately classifying URLs into True or False. This section elaborates on the working of every phase in detail. Fig. 1 displays the detailed methodology of Veracity Scanning Model (VSM) architecture, comprising of both iterations, diagrammatically.

1) Fetch healthcare-related web URLs: In this research, the authors have considered a document as a web page or URL text. Thus, to fetch the web URLs the authors have collected URLs from the Google search engine by using the set of keywords related to the healthcare domain. The list of 25 predefined keywords related to the healthcare domain along with their synonyms is maintained to get appropriate search results. The authors have collected 1000 URLs. Apart from these 1000 URLs, authors have fact-checked 200 URLs from healthcare based on expert opinions, existing valid datasets, and manual checking. This dataset of 200 URLs is used for Fact-checking.

2) Sentimental Bag-of-Words (s-BoW): In this phase, the textual contents of each URL are scrapped using a web scraper developed as a part of this research. The newly designed web scrapper can fetch only healthcare-related contents from the URL and remove non-healthcare-related contents. In the pre-processing stage, punctuations, single characters, stop words and duplicate data are removed thus reducing the size of the corpus and removing unwanted information appearing in the text. Further, a sentimental Bag-of-Words (s-BoW) related to the healthcare domain is developed. Initially, s-BoW contains manually identified and labeled sentimental words from the healthcare domain. This s-BoW evolves and grows as the model fetches and extracts new URLs incrementally.

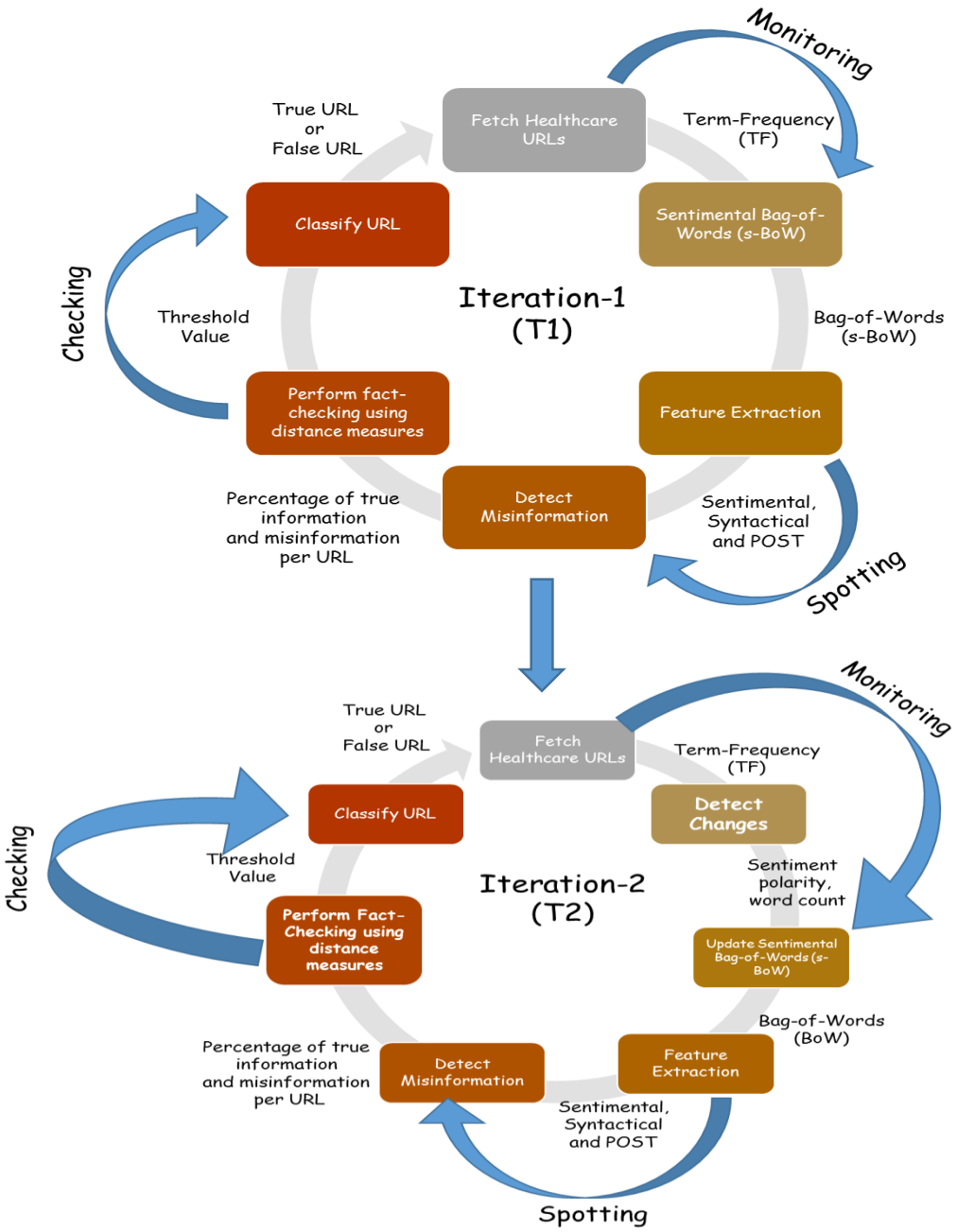


Fig. 1. Diagrammatic Representation of Methodology with the use of 'Veracity Scanning Model (VSM).

3) Feature extraction: In this phase, required features are extracted from the text using Term-Frequency (TF). The features extracted include a number of positive and negative words, count of words, nouns, pronouns, adjectives, and distances. The final list of features is the same as that used in the author's previous work.

4) Change detection: In the second and subsequent iterations, URLs are fetched to detect changes in the contents. The changes are detected based on the change in the word count, sentimental words, and sentence polarity. These changes are recorded and features are updated accordingly. Also, the sentimental Bag-of-Words is updated based on the newly arriving sentimental words. This helps to identify misinformation on new content.

5) Detecting misinformation, perform fact-checking, and classify URLs: This phase involves detecting the percentage of misinformation in URLs and categorizing them into True or False using a state-of-the-art classifier. The methodology to perform fact-checking is based on the author's previous work [1]. In this research, authors have devised a threshold-based fact-checking algorithm to perform fact-checking. The classification of URLs is based on the threshold value generated. An algorithm to compute the threshold value is shown in Fig. 2. Once the URL is fetched, the distance

between the incoming URL is computed with one of the URLs from the legitimate URL set using standard distance measure formulas of Jaccard Distance, Euclidean Distance, and Cosine Similarity [1]. The process is repeated for all the 2000 URLs at time T2. Further, a threshold value is computed by finding an average of all the distances of all the URLs. Thus, URLs are classified based on this threshold value. Apart from these classifications, Euclidean distance, Jaccard distance, and cosine similarity are used as a feature. The five state-of-the-art classifiers are used for classification viz. Logistic regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF).

Input: URL
Output: 1. Classification of URLs into Legitimate or Non-Legitimate
Algorithmic Steps:

1. Compute distance of URL w.r.t Legitimate URLs using Jaccard, Euclidean, and Cosine similarity measures
2. Compute threshold value T1 using distance with respect to Legitimate URLs separately for each distance measure.

$$T1 = \frac{\sum \ distance\ of\ all\ incoming\ URLs}{Total\ number\ of\ URLs}$$

3. if (distance $>$ T1)
4. Class= Non-Legitimate URLs
5. else
6. *Class= Legitimate URLs*

Fig. 2. A Threshold-based Fact-checking Algorithm for Classifying URLs.

IV. Experimental Results and Discussion

This section elaborates on the analysis of the results and performance of the model. The classification of URLs as Legitimate (True URLs) and Non-Legitimate (URLs with Misinformation) is performed using state-of-the-art classifiers viz. Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and also, threshold-based distance measure algorithms.

A. Performance Evaluation

The performance evaluation is measured through accuracy, precision, recall and F1-score and presented graphically respectively through Fig. 3 to Fig. 6 for document similarity measures on various classifiers. It can be seen that the RF Classifier outperformed the other 79.2% accuracy for the JD measure followed by LR Classifier 78.1% accuracy for JD Measure. The SVM model showed the least performance with an accuracy of 60.4% on the Cosine Similarity measure.

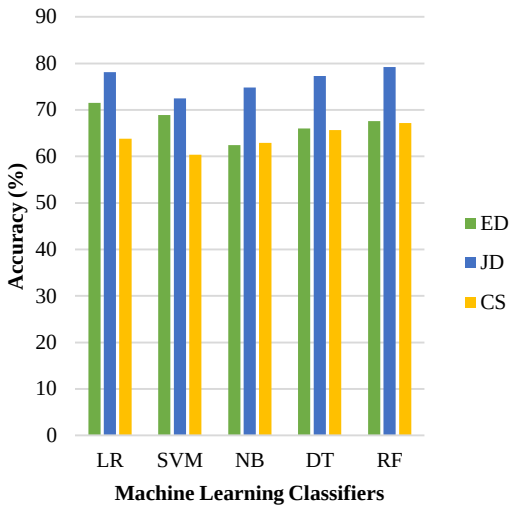


Fig. 3. Accuracy of the Proposed Model in Comparison with Standard Distance Measures.

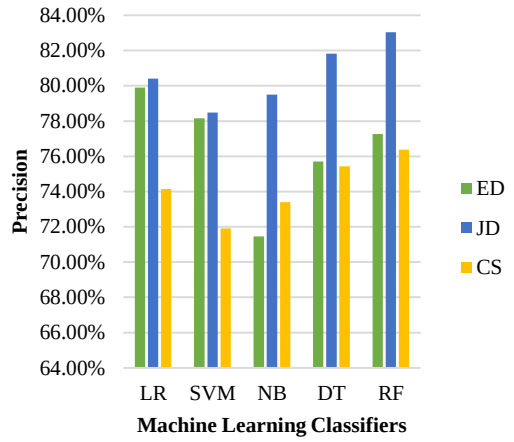


Fig. 4. The Precision of the Proposed Model in Comparison with Standard Distance Measures using Machine Learning Classifiers.

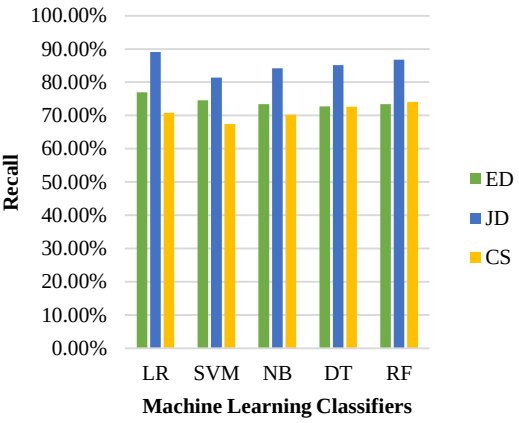


Fig. 5. Recall Matrix of the Proposed Model in Comparison with Standard Distance Measures using Machine Learning Classifiers.

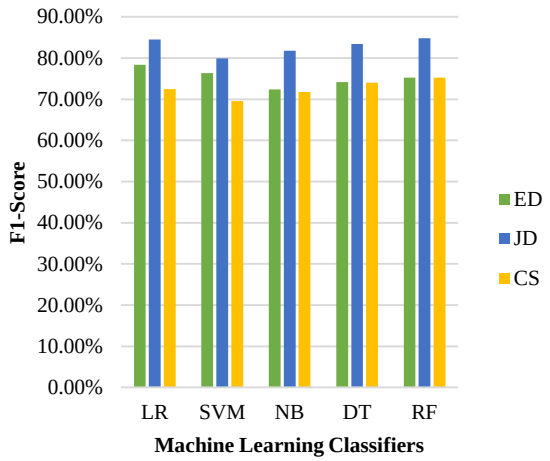


Fig. 6. F1-Score for the Proposed Model in Comparison with Standard Distance Measures.

Tables I to IV show the performance measures of accuracy, precision, recall, and F1-score, all in percentage and concerning distance measures for the five different classifiers, viz. LR, SVM, NB, DT and RF.

TABLE I. ACCURACY IN PERCENTAGE FOR THE PROPOSED MODEL FOR THE DISTANCE MEASURES

Accuracy			
	ED	**JD**	**CS**
LR	71.5	78.1	63.8
SVM	68.9	72.5	60.4
NB	62.4	74.8	62.9
DT	66	77.3	65.7
RF	67.6	79.2	67.2

TABLE II. PRECISION IN PERCENTAGE FOR THE PROPOSED MODEL FOR THE DISTANCE MEASURES

Precision			
	ED	**JD**	**CS**
LR	79.91%	80.40%	74.14%
SVM	78.16%	78.48%	71.90%
NB	71.45%	79.49%	73.41%
DT	75.70%	81.83%	75.43%
RF	77.27%	83.05%	76.38%

TABLE III. RECALL IN PERCENTAGE FOR THE PROPOSED MODEL FOR THE DISTANCE MEASURES

Recall			
	ED	**JD**	**CS**
LR	76.93%	89.14%	70.83%
SVM	74.55%	81.40%	67.41%
NB	73.36%	84.23%	70.24%
DT	72.77%	85.12%	72.62%
RF	73.36%	86.76%	74.11%

TABLE IV. F1-SCORE IN PERCENTAGE FOR THE PROPOSED MODEL FOR THE DISTANCE MEASURES

F1-Score			
	ED	**JD**	**CS**
LR	78.39%	84.54%	72.45%
SVM	76.31%	79.91%	69.59%
NB	72.39%	81.79%	71.79%
DT	74.20%	83.44%	74.00%
RF	75.27%	84.86%	75.23%

B. Analysis of the Proposed Model (VSM)

To evaluate the performance of VSM, the authors have analyzed the results of VSM at three different time intervals T1, T2, and T3. For time T1, the data of 2000 URLs were collected and analyzed to detect the percentage of misinformation in URLs and classify them into Legitimate or Non-Legitimate URLs after performing fact-checking. At time T2, once again the 2000 URLs are scrapped to detect any changes in the data. The changes are detected based on the count total number of words, sentimental words, and sentence polarity. Fig. 7 shows the number of URLs changed at time T2 and T3. It can be seen that at time T2 52 URLs have changed while at time T3 22 URLs have shown changes in data. Fig. 8 and Fig. 11 shows the change in percentage of misinformation due to change in incoming data at different time interval T1, T2, and T3. Thus, it can be seen from the figures that 50% of the URLs show a major increase in the percentage of misinformation at times T2 and T3. Another observation is that around 20% of the URLs showed a decrease in the percentage of misinformation. Also, 6 such URLs showed a change in data throughout the three iterations. The fluctuation in the percentage of these 6 URLs is shown in Fig. 10. It can be seen that 50% of URLs have increased in percentage of misinformation. Fig. 12, Fig. 13, Fig. 14 shows the confusion matrix of the VSM model for three-time intervals T1, T2, and T3. Fig. 9 displays the statistical analysis of the VSM model in terms of mean, mode, and standard deviation of legitimate and non-legitimate URLs. Hence, it has become a need of time to track the changes occurring in the data and update the model accordingly to detect and fact-check the newly changed data for misinformation increase or decrease. Incremental learning plays a key role to handle such a situation.

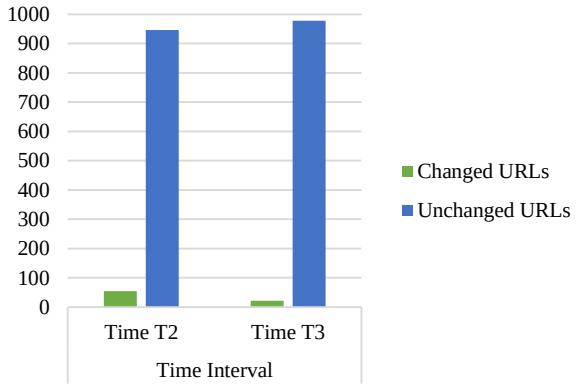


Fig. 7. Number of URLs changed at Time Interval T2 and T3.

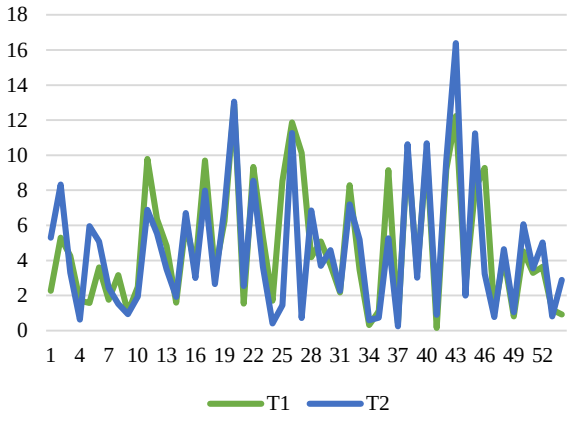


Fig. 8. URLs showing the change in Percentage of Misinformation at T1 and T2.

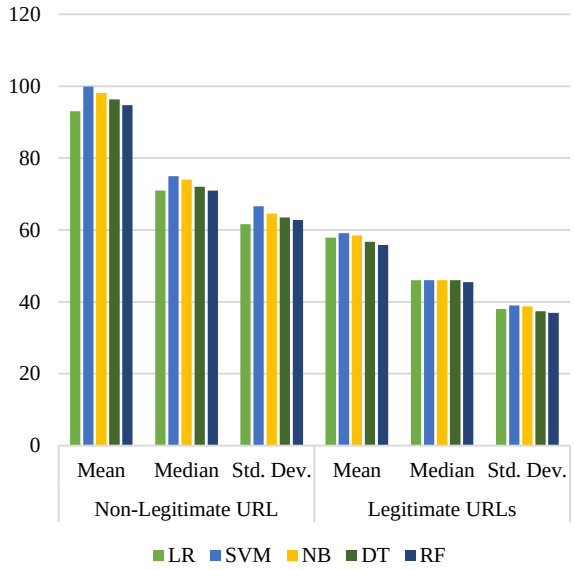


Fig. 9. Jaccard Distance Measure based on Mean, Mode, and Standard Deviation of the URLs.

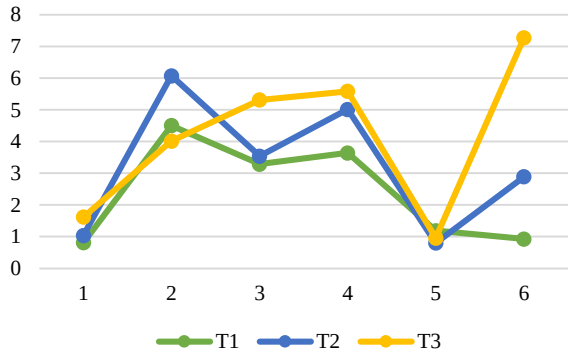


Fig. 10. URLs showing the change in Percentage of Misinformation in Three Iterations.

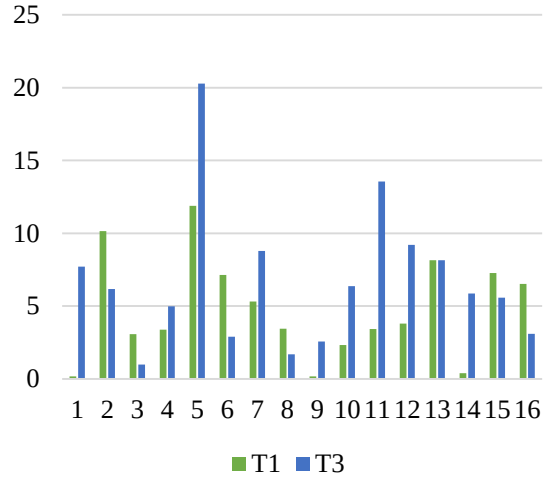


Fig. 11. URLs showing the change in Percentage of Misinformation at T1 and T3.

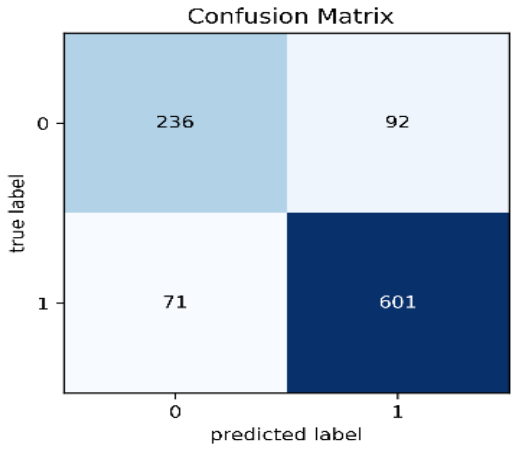


Fig. 12. Confusion Matrix at Time T1.

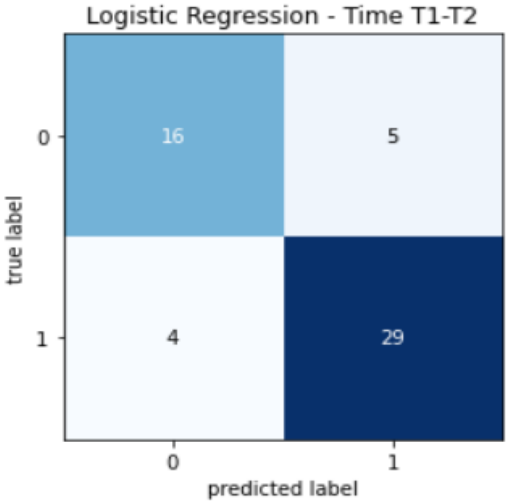


Fig. 13. Confusion Matrix at Time T2.

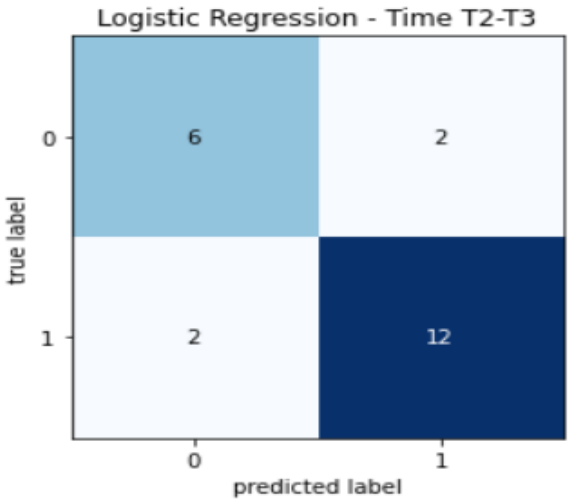


Fig. 14. Confusion Matrix at Time T3.

To evaluate the performance of the model on incremental data, 1000 URLs data was converted into five iterations. Each iteration has 200 URLs with a train and test split of 60% and 40% respectively. The threshold values for three distance measures viz. Jaccard, Euclidean, and Cosine are 0.177, 92.06, and 0.27 respectively. It can be seen from Fig. 15 that Jaccard distance showed max accuracy of 100% on test data, thus leading to model overfitting. However, in successive iterations, it showed an accuracy of about 90% to 93%. Euclidean distance measure showed maximum accuracy of 97.14% on training data followed by Jaccard distance and Cosine similarity with approx. 95% of accuracy. Also, the minimum accuracy for training and test data is 65% and 70% respectively using the Jaccard distance measure. Thus overall, the Euclidean distance measure performed well compared to others for both train and test data with an accuracy of 97.14% and 98.33%, respectively.

C. Comparison of VSM Model with Existing Technique

The proposed VSM model outperformed [22] in terms of accuracy. The work [22] showed the highest accuracy of 87.6% with random forest classifier using topic, linguistic, sentiment, and behavioral features while VSM showed an accuracy of 91.67%.

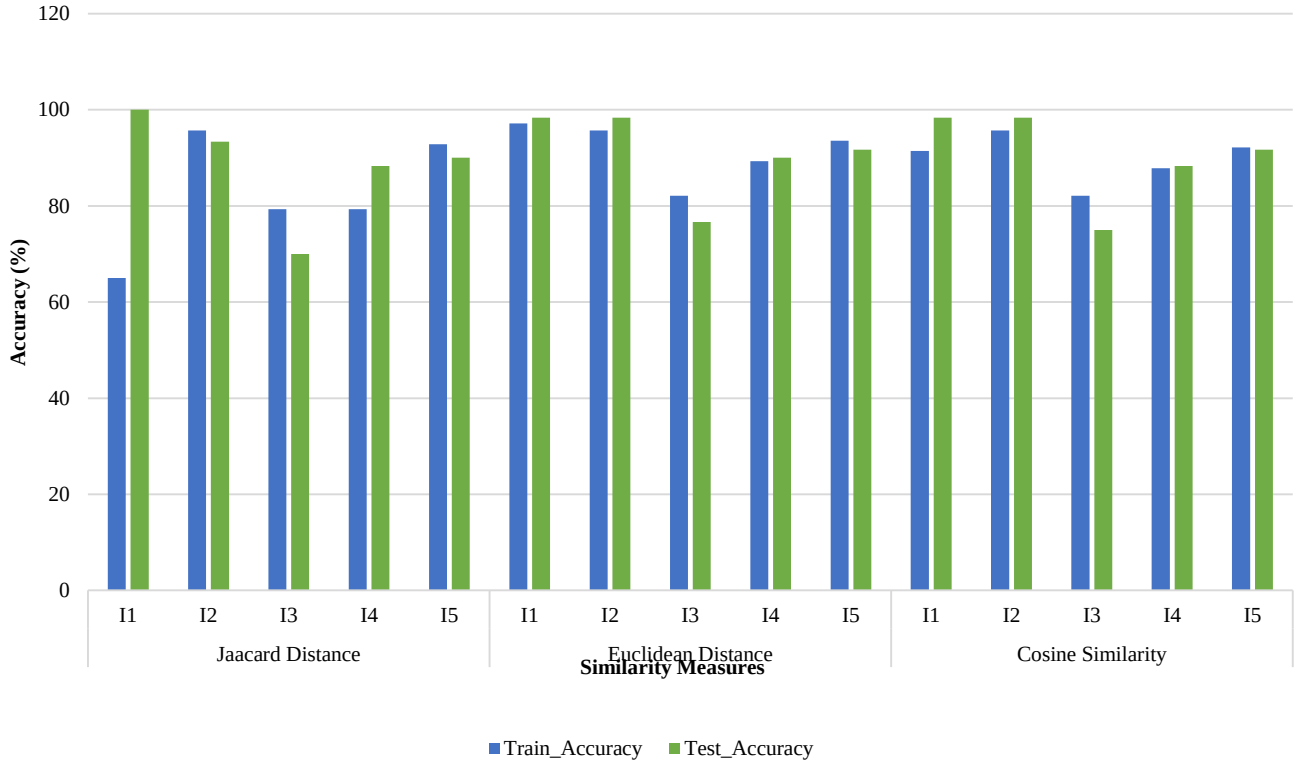


Fig. 15. Accuracy of Similarity Measures based Algorithms for Five Iterations of Incremental Data.

V. Conclusion and Future Enhancements

In this research authors have proposed a Veracity Scanning Model (VSM) using incremental learning, sentiment analysis, and document similarity approach. VSM overcomes the limitations of static models which fail to record changes at different time intervals. It was observed that URLs keep changing the contents with time and that fluctuates the percentage of misinformation in URLs. Therefore, to identify trustworthy URLs, especially in the healthcare domain there is a need for techniques like incremental learning to be adopted. The experimental results show that the Jaccard distance measure outperformed other distance measures with an accuracy of 79.2% with the Random Forest classifier, whereas the cosine similarity measure showed less performance of 60.4% accuracy with Support Vector Machine Classifier. Also, when implemented as an algorithm Euclidean distance

showed an accuracy of 97.14% and 98.33% respectively for train and test data.

In the future, the author wants to propose a new distance measure algorithm to classify URLs into legitimate and non-legitimate URLs and compare the performance with standard distance measures.

REFERENCES

[1] Y. Barve and J. R. Saini, "Healthcare Misinformation Detection and Fact-Checking : A Novel Approach", Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, pp. 295–303, 2021.

[2] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation", in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020, pp. 492–502, doi: 10.1145/3394486.3403092.

[3] J. J. McGowan and E. S. Berner, "Proposed Curricular Objectives to Teach Physicians Competence in Using the World Wide Web", Acad. Med., vol. 79, no. 3, pp. 236–240, 2004, doi: 10.1097/00001888-200403000-00007.

[4] S. Weinstein, "Internet for pathologists", Pathology, vol. 30, no. 4, pp. 364–368, 1998, doi: 10.1080/00313029800169646.

[5] X. Chen, F. Zhou, F. Zhang, and M. Bonsangue, "Catch me if you can: A participant-level rumor detection framework via fine-grained user representation learning", Inf. Process. Manag., vol. 58, no. 5, p. 102678, 2021, doi: 10.1016/j.ipm.2021.102678.

[6] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities", Expert Syst. Appl., vol. 153, 2020, doi: 10.1016/j.eswa.2019.112986.

[7] S. Shabani, Z. Charlesworth, M. Sokhn, and H. Schuldt, "SAMS: Human-in-the-loop approach to combat the sharing of digital misinformation", in CEUR Workshop Proceedings, 2021, vol. 2846, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104639416&partnerID=40&md5=325c6e737487df99352cf9c7de1ac333.

[8] B. Pecher, I. Srba, R. Moro, M. Tomlein, and M. Bielikova, "FireAnt: Claim-Based Medical Misinformation Detection and Monitoring", Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12461 LNAI, pp. 555–559, 2021, doi: 10.1007/978-3-030-67670-4_38.

[9] P. Lara-Navarra, H. Falciani, E. A. Sánchez-Pérez, and A. Ferrer-Sapena, "Information management in healthcare and environment: Towards an automatic system for fake news detection", Int. J. Environ. Res. Public Health, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031066.

[10] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 infodemic using explainable natural language processing models", Inf. Process. Manag., vol. 58, no. 4, p. 102569, 2021, doi: 10.1016/j.ipm.2021.102569.

[11] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking : A survey", no. August, pp. 1–21, 2021, doi: 10.1111/lnc3.12438.

[12] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar, "Fighting post-truth using natural language processing: A review and open challenges", Expert Syst. Appl., vol. 141, 2020, doi: 10.1016/j.eswa.2019.112943.

[13] S. Bhatt, N. Goenka, S. Kalra, and Y. Sharma, "Fake News Detection: Experiments and Approaches Beyond Linguistic Features", Lect. Notes Data Eng. Commun. Technol., vol. 71, pp. 113–128, 2022, doi: 10.1007/978-981-16-2937-2_9.

[14] W. Zang, P. Zhang, C. Zhou, and L. Guo, "Comparative study between incremental and ensemble learning on data streams: Case study", J. Big Data, vol. 1, no. 1, pp. 1–16, 2014, doi: 10.1186/2196-1115-1-5.

[15] P. Ksieniewicz, P. Zyblewski, M. Choraś, R. Kozik, A. Giełczyk, and M. Woźniak, "Fake News Detection from Data Streams", Proc. Int. Jt. Conf. Neural Networks, 2020, doi: 10.1109/IJCNN48605.2020.9207498.

[16] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, "False information detection in online content and its role in decision making: a systematic literature review", Soc. Netw. Anal. Min., vol. 9, no. 1, 2019, doi: 10.1007/s13278-019-0595-5.

[17] A. Chefrour, "Incremental supervised learning: algorithms and applications in pattern recognition", Evol. Intell., vol. 12, no. 2, pp. 97–112, 2019, doi: 10.1007/s12065-019-00203-y.

[18] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification", Expert Syst. Appl., vol. 169, 2021, doi: 10.1016/j.eswa.2020.114171.

[19] Y. Barve and P. Mulay, "Bibliometric Survey on Incremental Learning in Text Classification Algorithms for False Information Detection", Libr. Philos. Pract., vol. 2020, no. November, pp. 2388–2392, 2020.

[20] M. Umer, G. Dawson, and R. Polikar, "Targeted Forgetting and False Memory Formation in Continual Learners through Adversarial Backdoor Attacks", 2020, doi: 10.1109/IJCNN48605.2020.9206809.

[21] Z. Xu and H. Guo, "Using Text Mining to Compare Online Pro- and Anti-Vaccine Headlines: Word Usage, Sentiments, and Online Popularity", Commun. Stud., vol. 69, no. 1, pp. 103–122, 2018, doi: 10.1080/10510974.2017.1414068.

[22] Y. Zhao, J. Da, and J. Yan, "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches", Inf. Process. Manag., vol. 58, no. 1, 2021, doi: 10.1016/j.ipm.2020.102390.

[23] L. Kinkead, A. Allam, and M. Krauthammer, "Autodiscern: Rating the quality of online health information with hierarchical encoder attention-based neural networks", BMC Med. Inform. Decis. Mak., vol. 20, no. 1, 2020, doi: 10.1186/s12911-020-01131-z.

[24] H. Zhang, X. Xiao, and O. Hasegawa, "A load-balancing self-organizing incremental neural network", IEEE Trans. Neural Networks Learn. Syst., vol. 25, no. 6, pp. 1096–1105, 2014, doi: 10.1109/TNNLS.2013.2287884.

[25] Z. Chen, L. Huang, and Y. L. Murphey, "Incremental learning for Text document classification", in IEEE International Conference on Neural Networks - Conference Proceedings, 2007, pp. 2592–2597, doi: 10.1109/IJCNN.2007.4371367.

[26] Y. Liang, Z. Wen, Y. Tao, G. Li, and B. Guo, "Automatic security classification based on incremental learning and similarity comparison", in Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019, 2019, pp. 812–817, doi: 10.1109/ITAIC.2019.8785798.

[27] D. Wang and A. Al-Rubaie, "Incremental learning with partial-supervision based on hierarchical Dirichlet process and the application for document classification", Appl. Soft Comput. J., vol. 33, pp. 250–262, 2015, doi: 10.1016/j.asoc.2015.04.044.

[28] T. F. Rodrigues and P. M. Engel, "Probabilistic clustering and classification for textual data: An online and incremental approach", in Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014, 2014, pp. 288–293, doi: 10.1109/BRACIS.2014.59.

[29] T. Doan and J. Kalita, "Overcoming the challenge for text classification in the open world", in 2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017, 2017, doi: 10.1109/CCWC.2017.7868366.

Enhancing EFL Students' COCA-Induced Collocational Usage of Coronavirus: A Corpus-Driven Approach

Amir H.Y. Salama[1]
Department of English
College of Science & Humanities
Prince Sattam Bin Abdulaziz University
Alkharj, Saudi Arabia
Department of English, Faculty of Al-Alsun (Languages)
Kafr El-Sheikh University, Egypt

Waheed M. A. Altohami[2]
Department of English
College of Science & Humanities
Prince Sattam Bin Abdulaziz University
Alkharj, Saudi Arabia
Department of Foreign Languages, Faculty of Education
Mansoura University, Egypt

***Abstract*—The present study seeks to propose a novel pedagogical strategy for enhancing EFL students' collocational usage of the node 'coronavirus' as currently used in the Corpus of Contemporary American English (COCA) across its five genre-based sections, viz. TV/Movies, Blog, Web-General, Spoken, Fiction, Magazine, Newspaper, and Academic. Drawing on a corpus-driven approach, we conducted a pedagogical descriptive analysis of the 'coronavirus' top collocates generated by the COCA. The target collocates have been calculated by the Mutual Information (MI) of 3 or above and specified in terms of the four main lexical parts of speech of nouns, verbs, adjectives, and adverbs. The study has reached three main results. First, employing the COCA as a pedagogical corpus tool can enhance the collocational competence of EFL students should a corpus-driven approach be used descriptively in the classroom. Second, the two methodological stages of demonstration and praxis could facilitate the process of topical priority as a significant index of collocational usage and its thematic relevance. Third, more empirically, the naturally occurring collocates of the node 'coronavirus' have proven significant to the pedagogical situation of teaching the node's collocational meanings encoded in the syntactic categories of nouns, verbs, adjectives, and adverbs, e.g. *infection*, *cause*, *novel*, and *closely*, respectively.**

***Keywords*—COCA; collocations; coronavirus; corpus-driven approach; EFL learners; extended lexical units**

I. Introduction

Corpus-bereft research on collocations and their EFL usage can be said to remain captive of a good deal of misconceptions about the pedagogical nature of teaching and learning vocabulary at large. This type of research can readily be cited in support of the seemingly challenging claim stated above [1, 2, 3-6, 7]. Lamentably, drawing on limited sets of data, research of the sort has offered results about EFL collocational usage that are insensitive to balanced genres in the target learning language, not least because the various patterns of co-occurring words have been conspicuously absent from the analysis of lexical items with collocates used in various domains of human experience.

In an attempt to tout a practical solution to the foregoing problem, we propose to utilize the Corpus of Contemporary American English, commonly known and cited as COCA [8]. The node selected as a model for corpus collocational analysis is the lexical item 'coronavirus' as a currently globally used term in multifarious English-language genres. Towards the collocational analysis of 'coronavirus' in the COCA, we adopt a corpus-driven approach of the extended lexical unit [9, 10] as a methodological tool whereby the following research question can be addressed: How can the COCA be utilized in enhancing EFL students' collocational usage of 'coronavirus'?

Indeed, the question raised above should highlight the significance of the present study as a highly pedagogical and empirical medium for EFL teachers as a community of practice. It is through such a medium that innovative corpus-driven methods can be used for easing EFL students' comprehension of the collocational meanings associated with lexemes of wide-scale thematic relevance and topical interest within the same community of practice. The practical example of the lexeme 'coronavirus', alongside its potential collocates, is claimed to be a practically good and productive site of such relevance and interest. On a more general note, the corpus-driven approach adopted in the current study is likely to secure the ligature between the use of computationally and linguistically tagged corpora such as COCA and the pedagogical applications of teaching one of the most problematic areas in the study of English as an FL – collocational meaning.

In keeping with the corpus-driven approach, therefore, we posit an EFL hypothesis that is subject to empirical validation; the hypothesis can be formulated as such: Applying a corpus-driven approach to the COCA can aid in enhancing the EFL students' collocational usage of the node 'coronavirus'. Addressing the question raised above, and thus (dis-)proving the foregoing hypothesis, the upcoming structure of the paper unfolds in the following way. First, Section 2 presents a review of the literature relevant to the principal point of research. Second, Section 3 elucidates the study's corpus-driven approach as a theoretical framework for the corpus data analysis. Third, Section 4 outlines the current research methodology with a focus on the corpus data (COCA) and the procedure of analysis. Finally, Section 5 concludes the study

by offering a discussion of the main findings and a prospect for future research.

II. Literature Review

Language learners' use of collocations received notable scholarly linguistic interest with the focus of exploring their patterns, distribution and frequency, identifying common errors in collocational usage, and experimenting research methods in collocation research. Findings of related studies should ideally be reinforced with investigating the collocational behaviour of pedagogic search terms, identifying their lexico-grammatical patterns, enhancing and testing collocational competence, and gaining insights regarding collocation learning and teaching. The methods employed in collocation research can be generally divided into two directions. The first direction explores the production of collocations by means of large learner corpora [11-18]. The second direction targets collocations collected from questionnaires, interviews, and tests, especially translation [19-22].

Based on collocation sets extracted from learners' essays, Granger [13] compared the way native and non-native learners of English used collocations structured as intensifiers + adjectives. She found that non-native students used atypical word combinations marked as unacceptable by non-native students. Likewise, Howarth [14] focused on the verb-object collocations used by native and non-native learners in different written modes. Findings showed that both native and non-native learners produced non-standard – especially restricted – collocations. Similar findings were reported in Nesselhauf [46] who affirmed that non-native learners' errors in using combinations were distributed over a continuum ranging from free combinations to idioms.

In a similar vein, Durrant and Schmitt [12] compared the native and non-native use of highly frequent collocations used in two parallel corpora composed of students' writing assignments in pre-sessional and in-sessional courses in the UK and Turkey. They targeted manually extracted adjacent modifier-noun word pairs claimed to be particularly common. Using the association measures of Mutual Information (MI) score and T-score, the frequency and strength of such collocations were compared to the same collocations used in BNC as a reference corpus. Findings showed that unlike non-natives, native learners tended to use more low-frequent combinations out of conservatism while writing long essays. Also, non-natives significantly overused strong collocations, but they showed a significant preference to the use of particular combinations.

Altenberg and Granger [11] applied a corpus-based approach, by means of WordSmith Tools, for exploring EFL French/Swedish learners' use of highly frequent collocations based on the verb 'make'. An authentic learner corpus was compared with a native-speaker corpus, namely, the Louvain Corpus of Native English Essays (LOCNESS). Findings highlighted that even advanced learners misused collocations. Although eight uses of 'make'-based collocations have been identified, learners underused delexical (e.g. 'make a decision') and causative (e.g. 'make something possible') structures. A similar approach was followed by Laufer and Waldman [15] who compared learner (the Israeli Learner Corpus of Written English, ILCoWE) and native speaker (LOCNESS) corpora regarding the frequency and correctness of verb-noun collocations. Findings showed that unlike native speakers, non-native learners used fewer collocations. Furthermore, even more advanced learners misused collocations. Also, Paquot and Granger [18] explored the use of English formulaic language in learner corpora, including collocations, phrasal verbs, compounds, idioms, speech formulae, etc. Findings affirmed the relative negative impact of L1 on learners' use of formulaic language regardless of their proficiency level.

Li and Schmitt [23] were concerned with how far L2 learners' collocational competence develops over a year of training on the usage of collocations in an academic writing course in an MA English language teaching program. The reported findings showed no statistically significant development in learners' knowledge of collocations as they tended to overuse specific collocations. Certain errors remained unchanged as learners relied heavily on creativity rather than following lexical patterning. Similarly, Nguyen and Webb [17] explored Vietnamese EFL learners' knowledge of collocations at different frequency levels, the correlation between knowledge of collocations and single-word items, and the predictors of receptive knowledge of collocation. Findings affirmed the positive correlation between knowledge of collocations and single-word items. Also, the major predictors of receptive knowledge – and accordingly the learnability – of collocations included node word frequency, collocation frequency, mutual information score, collocation congruency, and part of speech.

Bahns and Eldaw [19] focused on testing the collocational competence of advanced EFL German learners' by means of translation activities and a cloze test. Findings showed that students sought to paraphrase collocations. Even though some collocations were successfully paraphrased, most paraphrases were unacceptable. Therefore, paraphrasable collocations should not be given prominence in English language teaching. Unlike Bahns and Eldaw [19], Farghal and Obiedat [21] compared the collocational competence of two groups. While the first group included junior and senior Jordanian students at Yarmouk University, the second included English language teachers. Towards this objective, two questionnaires have been administered in the form of fill-in-the-blank and translation tasks. Findings demonstrated that learners' deficiency in using collocations forced them to use lexical simplification strategies, e.g. synonymy, paraphrasing, avoidance, and transfer.

Biskup [20] explored the challenges that faced Polish and German university learners in translating lexical collocations into English. Such translations were then assessed by native speakers of English in terms of acceptability and equivalence. Findings affirmed that both Polish and German students had translational errors. However, while German learners' errors were due to similarity in form, Polish learners' errors were ascribed largely to extending the meaning of L1 collocations to L2. Similarly, Hasselgren [22] explored Norwegian learners' awareness of English collocations during translation tasks. He affirmed that learners' misunderstanding and poor

knowledge of collocations led them to rely on literal translation creating what he describes as "collocational dissonance." That is, though the emerging collocations were grammatically sound, they were not native-like.

Given the Corpus of Contemporary American English (COCA) as the target corpus of present study, several studies have affirmed its efficacy in enriching students' collocational use especially in writing [24-29]. Hu [25] explored the challenge that near-synonyms impose on the learnability and use of collocations for EFL students. The target synonymous adjective pairs were 'initial/preliminary', 'following/ subsequent', and 'sufficient/adequate'. Whilst such pairs were used interchangeably in isolation, findings showed that these collocates designate different prosodies (positive, neutral, and negative) in academic discourse with diverse attitudinal and evaluative meanings.

Mansour [27] sought to foster L2 students' use of collocations for improving their writing competence and translation performance through getting them to use COCA effectively. Using the COCA's list display and collocates display options, students have shown significant development in using collocations after receiving the proper training. Likewise, following quasi-experimental research design, Kartal and Yangineksi [26] explored how EFL students learn and produce verb-noun collocations. Hence, experimental and control groups were created, and a collocation knowledge test was administered before and after training students to use collocations through the COCA concordance tool. Findings showed statistically significant differences between the experimental and control groups in terms of the production of collocations. Yet, no significant differences have been noted regarding their collocational knowledge. Similarly, Fang, Ma and Yan [24] explored the way corpus-based training on data-driven learning activities could improve Chinese secondary school students' writing performance and vocabulary competence in IELTS, including the use of collocations. Students were trained to search two main corpora: COCA and Word and Phrase Concordance. Towards fulfilling this main objective, pre-writing and post-writing tests were used. Findings affirmed that students' performance in word selection significantly improved as the frequency of collocational errors decreased.

Oktavianti and Sarage [28] studied the frequent and strong collocates of the adjectives 'great' and 'good' in a corpus compiled from Indonesian EFL textbooks and compared them with those used in COCA. Based on the MI score of collocates, both corpora were similar regarding the verb + adjective structure (e.g. 'look great/good'). However, considerable mismatches were reported regarding the adverb + adjective structure (e.g. 'pretty good' and 'unpredictably great'), and prominent collocations following the structure of adjective + noun (e.g. 'great deal' and 'good idea') were markedly absent. Accordingly, textbooks were recommended to be re-examined to render the presented collocations more authentic. Relatedly, Wu [29] investigated how Taiwanese students studying English used the COCA, in an essay writing course, for discovering the collocational patterns of thirty near-synonymous change-of-state verbs. Towards this objective, mixed methods were used including pre-, post-, and delayed post-tests, video files of corpus consultation, a questionnaire, and interviews. Findings showed that although students had some challenges in using the COCA in correcting their miscollocations while drafting their essays, their performance in using collocations improved and such improvement lasted for a considerable time as affirmed by the delayed post-test results.

In view of the foregoing literature review, there seems to be a problematic paucity of corpus-driven investigations of collocations that reflect globally thematic significance and relevance to EFL students/learners in general. Indeed, the collocational use of lexemes whose magnitude of topical saliency and eventfulness is imposing in various semantic domains of expression can be crucial to EFL learners/students at the pedagogical level. One such exemplar is the globally used search-term lemma 'coronavirus'; and since the term has become a de facto topical attraction in classrooms, either in translation or in writing, there needs to be a particular concern with and focus on the lemma's collocational usage. This should be especially so at the syntactic level of different parts of speech in genre-balanced corpora wherein collocational usage is likely to be conducive to enhancing competence and developing idiomatic expression. As a corollary of this research gap, the present study attempts to investigate the collocational usage of the lemma 'coronavirus' in the COCA in a bid to enhancing the EFL competence of using the currently widely used lemma, and thereby improving the students'/learners' performance when it comes to using the word-forms associated with this lemma in various pedagogic settings.

III. Theoretical Framework

The lexical meaning of a word is often determined in light of the words that syntagmatically co-occur with it. Such words that tend to hang out together as ready-made chunks came to be known as 'collocations'. A collocation is commonly viewed as a multi-word formulaic unit (lexical bundle) just like idioms (e.g. 'back to square one'), proverbs (e.g. 'let's make hay while the sun shines'), functional expressions (e.g. 'excuse me'), fillers (e.g. 'kind of'), and standardized phrases (e.g. 'there is a growing body of evidence that') [30]. Cruse [31] defines collocations as "sequences of lexical items which habitually co-occur" (p. 40). Indeed, Nattinger and DeCarrico [4] define collocations as "strings of words that seem to have certain 'mutual expectancy', or a greater-than-chance likelihood that they will co-occur in any text" (p. 21). Hoey [32] affirms that high frequency is the most salient principle marking the behaviour of collocations in a language. He assumes that collocations refer to "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (p. 7) [32].

Despite its classification as a unit of formulaic language, a collocation, unlike an idiom, is a compositional phraseme (i.e. a phraseological unit) since its meaning is relatively transparent, i.e. it can – but not necessarily – be inferred from the meaning of its individual parts. Structurally speaking, a collocation is substitutable even when synonyms or near-synonyms are applied; for instance, 'strong/powerful argument' is a recurrent collocation, while the collocations

'strong car' and 'powerful tea' are awkward. Also, while a combination such as 'good fortune' is recurrent, the combination 'nice fortune' is semantically unacceptable. Furthermore, unlike free combinations, collocations are somehow considered "grammar in terms of vocabulary" (p. 216) [33], i.e. their co-occurrence always adheres to a set of grammatical principles.

Collocations still, however, could be distributed over a phraseological continuum [34] (Fig. 1) ranging from free combinations (e.g. 'want a car'), to restricted colocations (e.g. 'hold a discussion'), and finally to frozen idioms (e.g. 'sweeten the pill') [14]. The items in free combination are easily replaceable in terms of grammar. Yet, unlike frozen idioms, the meaning of collocations is much more transparent.

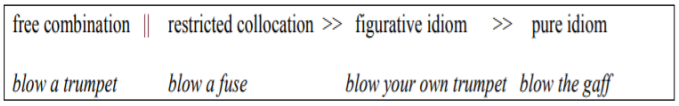


Fig. 1. Cowie's Phraseological Continuum [34].

Gledhill [35] posits that the term 'collocation' tend to signify different notions as far as different perspectives are adopted. First, from a statistical/textual perspective, a collocation signifies a node and its collocates recurrent in a text, i.e. a particular lexical item frequently accompanies another lexical item due to constraints of usage. Therefore, the collocational patterns of a particular phrase are triggered by other phrases at a distance. Second, from a semantic/syntactic level, collocations are approached in terms of lexical combinability, i.e. they are regarded as recurrent, restricted composite units of meaning arranged in particular grammatical sequences, taking into consideration that such sequences are inseparable from their propositional meaning. This collocational restriction means that the meaning of an individual word in specific two-word collocations is restricted or confined to such collocation [36]. For instance, the word 'white' in 'white coffee', 'white noise', and 'white wine' has different senses. Finally, from a discoursal/rhetorical perspective, collocations are assigned diverse pragmatic functions across discourses such as marking topics (e.g. 'let's look at'), shifting topics (e.g. 'ok now'), summarizing (e.g. 'so then'), relating (e.g. 'it has to do with'), and qualifying (e.g. 'the catch is that').

Insofar as the classification of collocations is concerned, two approaches can be enlisted: the phraseological and the distributional [17, 37]. While the phraseological approach focuses on the semantic relation among the words forming the collocation and the non-compositionality of their meaning, the distributional approach focuses on the frequency of a collocation in a corpus or corpora. In view of this demarcation and based on the items forming collocations, collocations are classified into lexical and grammatical collocations. Grammatical collocations are more frequent in English, and they are lexicalized as single units with formulaic meanings. Bahns [38] affirms that grammatical collocations take the form of a noun, an adjective, or a verb followed by either a particle, an infinitive, or a clause, e.g. 'by accident', 'angry at', 'afraid that', and 'adhere to'. Unlike grammatical collocations, lexical collocations have no grammatical elements as they are composed of open-class lexical items [39]. They are structured as noun + noun (e.g. 'ceasefire agreement'), adjective + noun (e.g. 'strong tea'), noun + verb (e.g. 'results showed'), verb + noun (e.g. 'make a mistake'), verb + adverb (e.g. 'walk slowly'), and adverb + adjective (e.g. 'amazingly gorgeous'). Furthermore, clusters of lexical collocations are claimed to share a similar semantic prosody [40]. Yet, based on the frequency of collocations, Hill [41] divides collocations into four types: weak collocations (e.g. 'red wine'), medium-strength collocations (e.g. 'Sun reader'), strong collocations (e.g. 'rancid butter'), and unique collocations (e.g. 'leg room').

With regard to collocational use, a set of parameters have been proposed. One crucial parameter of collocational use is high frequency, i.e. highly frequent word combinations are systematically classified as collocations. Another important parameter is that of word association strength as specific words tend to co-occur biasedly [42]. In this regard, diverse statistical measures could be employed such as Mutual Information (MI), T-Score, and Z-Score. A third parameter is that of substitutability, i.e. how far an item in a collocation could be substituted by a synonym or a near synonym. As an integrative part of any language, collocations are processed in the mind of language user in two different ways: analytic processing and holistic processing [43]. On the one hand, in analytic processing, the lexical items and grammatical patterns of word combinations are computed and then their meanings are retrieved by assembling each item's meaning. This kind of processing occurs at a slower speed with much processing load. Holistic processing, on the other hand, is conducted at a faster speech with less processing load as word combinations are memorized as units (prefabricated forms) whose meanings are relatively difficult to be retrieved from the meanings of their individual parts, such as phrasal verbs and compounds. Still, collocation processing might occur in a parallel mode [43].

Indeed, the surge of publications in collocation teaching and learning is by all accounts an index of the significance of collocation competence, which is in turn crucial to language proficiency. For instance, Gledhill [35] asserts that "it is impossible for a writer to be fluent without a thorough knowledge of the phraseology of the particular field he or she is writing in" (p. 1). Equally important, Hill and Lewis [44] regard collocation use as "one of the most powerful forces in making language coherent, fluent, comprehensible, and predictable" (p. 1). Similarly, in any text, collocations are claimed to have "a cohesive force" [45]. Additionally, Fillmore, Kay, and O'Connor [46] point out that collocations should be integrated in language learning since they are culturally salient.

Hill [41] affirms that collocations represent "the most powerful force in the creation and comprehension of all naturally occurring text" (p. 49). In the context of English Language Teaching (ELT), it has been largely claimed that the accurate use of collocation is an essential component of communicative competence [47] and an indicator of proficiency as collocational knowledge allows native-like language use [48]. That is why the misuse of collocations is envisaged as "a major indicator of foreignness" (p. 232) [38].

That is, most of the collocational errors are experienced by non-native speakers usually due to lack of lexical proficiency. In this regard, Nation [3] explains that less proficient learners tend to "encode words in memory on the basis of sound and spelling rather than by association meaning" (p.3). Similarly, Laufer [49] and Erman, Lundell and Lewis [50] affirm that collocations, among other formulaic units are linked to native speakers' fluent and natural language production as well as linguistic diversity.

Crucially, a methodological distinction is always made between corpus-based and corpus-driven approaches. The corpus-based approach (CBA) targets previously identified linguistic features and constructs as well as patterns of variation and use. In other words, the corpus would support intuitive knowledge, confirm linguistic pre-set assumptions, and provide illustrative examples. Meanwhile, the corpus-driven approach (CDA) aims at exploiting the potential of corpora for the identification of recurrent linguistic categories and patterns emerging in context not fully recognized before [51-52]. Furthermore, CBA starts with no prior assumption, and all conclusions are usually reached relying on corpus observations. The corpus-driven approach has been largely used in the analysis of multi-word sequences known as 'lexical bundles' including idioms, proverbs, and collocations. The target is always their frequency, and distribution, usually followed by an analysis of emerging patterns and functional characteristics.

Indeed, many factors have been reported to affect the learnability of collocations. One of these factors is in the semantic complexity of a collocation. Regarding semantic complexity, collocations could be distributed over a continuum from total transparency to opacity. Figuring out the meaning of a collocation depends on the language user's familiarity with the individual words forming the collocation. Accordingly, it is largely claimed that learners are expected to spot the meaning of free combinations (e.g. 'pay money') more than restricted (e.g. 'pay attention'), and idiomatic collocations (e.g. 'pay lip service') [15, 53]. Furthermore, collocational congruency is also claimed to affect collocational usage as EFL students are reported to make more errors and react more slowly to incongruent collocations than congruent collocations (p. 647) [54]. Specifically in restricted collocations, L2 learners are reported to make "overliberal assumptions about the collocational equivalence of semantically similar items" (p. 202) [48]. That is, they tend to be able to produce atypical word combinations using items with similar meanings, e.g. 'plastic operation' instead of 'plastic surgery'. The reason is that they perceive lexical items individually rather than in combination, and therefore they strategically tend to simplify the lexemes form collocations through the use of synonyms, paraphrasing, and transferring L1 items to L2 through literal translation [21]. Generally, L1 interference is largely claimed to produce many of the errors in collocational usage, even on the part of advanced learners [55]. Such words that learners learnt at early stages and tend to cling to them even after training came to be known as "lexical teddy bears" [22].

Indeed, the introduction of corpus tools and techniques formed a turning point in phraseology studies in general and the study of formulaic units (including collocations) in particular. L2 research benefited greatly from such tools and techniques which offered more comprehensive empirical techniques for building, analyzing, and comparing corpora, thereby allowing the exploration of authentic language as practised by language learners compared to native speakers. This line of research is referred to as 'data-driven learning' (DDL) [9, 56, 57]. In DDL, language is viewed as data and the main objective of DDL tasks is to lead learners to identify patterns and uses of language by means of corpus tools, and thereby developing their autonomy. Software packages such as WordSmith, ConcGram, AntConc, etc. offer tools for calculating frequencies of words and their token/type ratio, extracting concordance lines with the target key words, featuring their various co-texts. As mentioned in the literature review section, corpora – taking native-speaker corpora as the norm – have been employed in investigating the typology of collocations as well as collocational underuse, overuse, and misuse.

IV. Methodology

A. Data

The present study is geared towards eliciting the highly frequent collocates used by native speakers of General American English (GA) when covering information of different types on the outbreak and progression of 'coronavirus' (or more technically, COVID-19) as represented in the American Corpus of Contemporary American English (COCA) [10]. The COCA has been selected in this study for a host of reasons. First, updated in 2021, the COCA (available at https://www.english-corpora.org/coca/) contains more than one billion words of data distributed over 485,202 texts, and therefore it is regarded as the most widely used, freely available corpus worldwide. Secondly, it is genre-balanced as it offers data covering a wide range of spoken and written, formal and informal genres. These genres are web genres and blog, newspapers, magazines, spoken, academic, fiction, and TV/movies. Thirdly, and finally, the user-friendly interface of COCA allows getting information about – and comparing – the frequency, currency, time span, and prosody of words, phrases, and grammatical constructions across diverse genres. Besides, the COCA offers information on definitions, keyness, related topics, collocates, clusters, lemmas, synonyms, and customized word lists.

B. Procedure

The methodological procedure adopted in the present study is a two-stage process of addressing the primary research question of how the COCA can be utilized in enhancing EFL students' collocational usage of 'coronavirus' (see Fig. 2). The first stage is concerned with setting the pedagogical scene; it amounted to a demonstration of how the COCA's interface can potentially be utilized in terms of its available POS syntactic tagging and the sort/limit function as well as the frequency cut-off point and hits specified. In respect of the second stage of the procedure, the lexical item 'coronavirus' has been presented with its assigned part-of-speech collocates with the automatic generation calculated by an MI score of 3 or above and the collocability default range ±4. At this stage, too, the frequency distribution of

'coronavirus' over the COCA's genre-based sections was calculated based on the generated collocates themselves, and thereafter the topical priority associated with 'coronavirus' collocates in COCA was automatically retrieved in relation to the extracted collocational pairs.

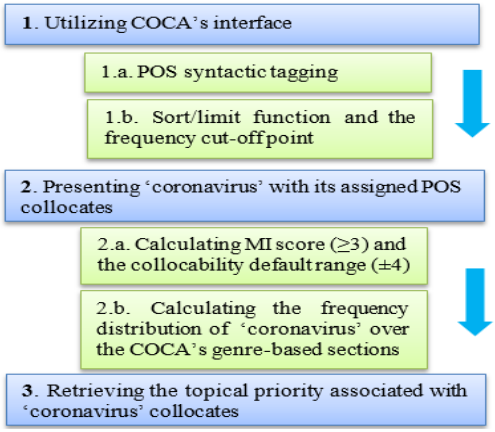


Fig. 2. The Procedure of Analysis.

V. Data Analysis and Discussion

The present section of analysis is divided into two stages. The first stage is dedicated to the demonstration of the COCA's interface as a way of setting the pedagogical scene; the second stage is a proposed EFL pedagogical praxis whereby the automatic generation of 'coronavirus' collocates and their relevant topical priority across the COCA's different genre-specific sections.

A. Setting the Pedagogical Scene: The COCA in Focus

The current stage of analysis can best be described as an EFL demonstration of the COCA's interface. As exhibited in Fig. 3, the lexical item 'coronavirus' has been entered into the COCA's search box with the particular POS tagging noun. ALL, which restricts the search hits to coronavirus as an exclusively nominal form. Also, demonstrably, the different COCA sections are presented for a potential selection, whereby an EFL teacher can decidedly opt for a genre-based search domain for 'coronavirus', say, TV/MOVIES or FICTION; and, perhaps, the teacher can interestingly compare such sections.

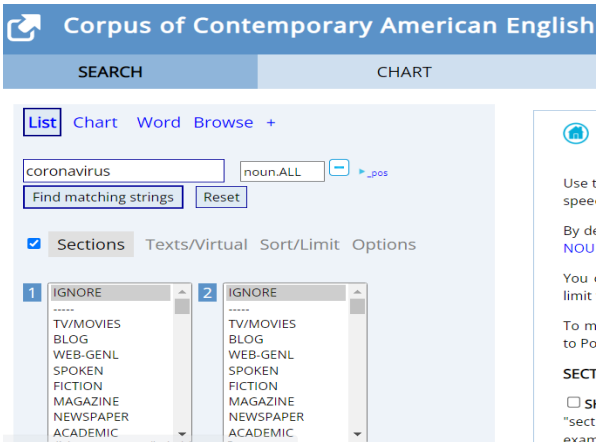


Fig. 3. The COCA's Genre-based Sections and POS Tagging.

Fig. 4. The COCA's Sort/Limit Function for Lemma Search.

Moving to Fig. 4, EFL students can be trained in how to use the COCA's Sort/Limit function for lemma search. The teacher is just supposed to employ this function as a way of specifying the frequency cut-off point of 20. The function is crucial since it facilitates the pedagogic situation by rendering the search process manageable enough to the students, let alone the fact that the same function generates the highly frequent occurrence of the lexical item as a lemma.

There are yet other COCA-built functions for lemma search as shown in Fig. 5, where other options are displayed. At this point, the teacher should ideally keep the students focused on the number of hits germane to 'coronavirus' (100 times) and the KWIC scope allowed (by default 200) as well as the featured raw frequency; these more options can further be used to facilitate the process of searching for the significant instances of the lemma 'coronavirus'. Of course, thus far, we have not touched upon the actual distribution the lemma ('coronavirus') over the genre-based sections – which is so pedagogically crucial to the recognition of vitiations in use.

Fig. 5. The COCA's Other Significant Functions for Limiting Lemma Search.

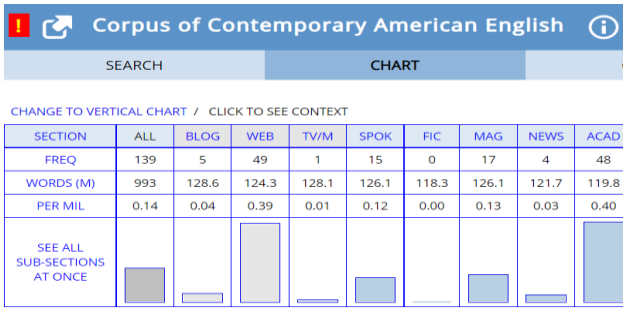


Fig. 6. Frequency Distribution of 'Coronavirus' Over the COCA's Genre-based Sections.

Coming to Fig. 6, the teacher can be said to be able to prepare his/her students for the recognition of the variations in lemma use referred to above, with a closer eye on frequency distribution. Thus, on closer inspection, students will readily observe that 'coronavirus' is most frequently used in the section WEB PAGES next to which in frequency is ACADEMIC. Perhaps, this may be ascribed to the fact that the Internet as an electronic medium has consistently demonstrated global-scale impact due to "its intensity of use" (p. 5) [58]. The frequency associated with the COCA's section ACADEMIC can be explained on the grounds that the topical nature of 'coronavirus' is scientific in the first place. Also, the low frequency of 'coronavirus' in the sections of FICTION (0 frequency), TV/M (1 time), NEWS (4 times), and BLOG (5 times) can be explained against the background genre nature. A good teacher, we argue, ought to think of the widespread use of a given lemma in a certain genre, and here there lies the rub: compared to the WEB PAGES and ACADEMIC, the rest of the low-frequency genres lack in sub-genres. Thus, when it comes to searching for a currently global term such as 'coronavirus', students should be directed to sub-genre-composed genres.

As the current EFL demonstration proceeds, we need to draw teachers' attention to the fact that in order for students to gain the utmost pedagogical benefits out of the COCA's interface, the frequency of 'coronavirus' or of any other search term is far from enough; there needs to be an investigation of the patterns of use associated with 'coronavirus', or again with any comparable term. At this point of analysis, therefore, collocational usage of 'coronavirus' is presented as a pedagogical praxis in the coming sub-section.

B. Proposing a Pedagogical Praxis: The COCA-Driven Collocates of 'Coronavirus'

EFL students' competence for the collocational usage of 'coronavirus' – and indeed any other lexical item – can be well enhanced should the teacher make a point of searching for the collocates strongly co-occurring with the node word 'coronavirus'. As exhibited in Fig. 7, this is feasible via the COCA's interface by clicking the 'collocates' icon after specifying the POS tagging of target collocates. The figure features the tagging adj. ALL of such collocates and offers the default range ±4, i.e. four collocates to the right and/or the left of the search term as the topmost span (4:4 collocates). Further, crucially, the teacher needs to present students with the collocation measurement of Mutual Information (MI of 3 or above).

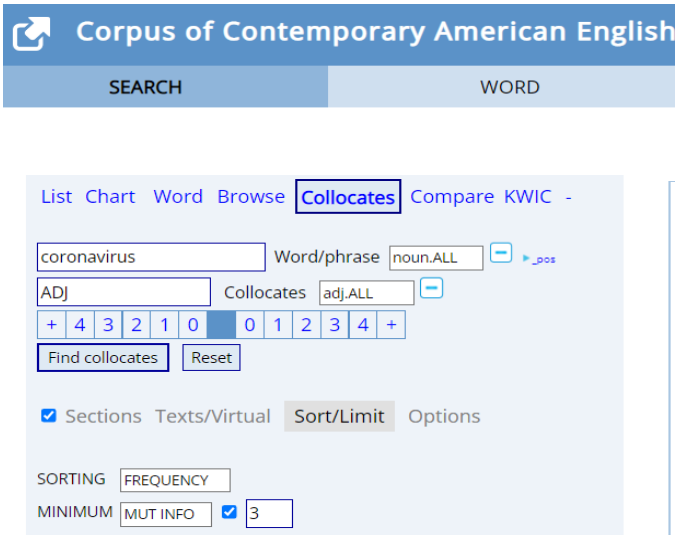


Fig. 7. Searching for the Collocates of 'Coronavirus' on COCA's Interface.

Moving to the next practical step in EFL class demands a pedagogically direct engagement with the top collocates of 'coronavirus' in terms of their lexical parts of speech, i.e. nouns, adjectives, verbs, and adverbs. Indeed, this is one of the most useful tools of the COCA's interface. As shown in Fig. 8, the tool is corpus-driven, in that it provides the top part-of-speech collocates based on their frequency of co-occurrence with the node 'coronavirus' in varying degrees of highlighting. Thus, the order of noun collocates are infection, probability, syndrome, sar, virus, and ferret; the adjective collocates are novel, respiratory, and new; the verb collocates are cause, identify, confirm, and ferret; the adverb collocates are closely, newly, previously, meanwhile, and widely. Given such a corpus-driven set, students should be equipped with a whole profile of the main collocates in their various lexical parts of speech; at this point, the teacher is required to engage with the teaching situation by asking students to use collocations in different lexico-grammatical patterns, or to emulate certain native-like usages of comparable patterns.

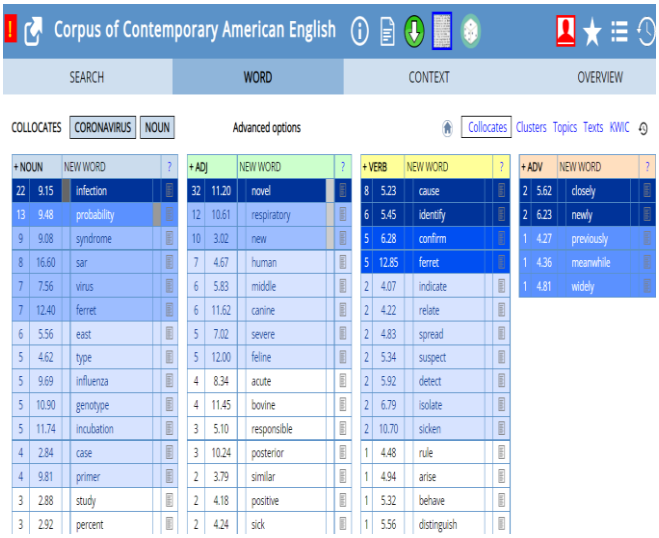


Fig. 8. Top Collocates of 'Coronavirus' and their Lexical Parts of Speech in COCA.

Fig. 9. Topical Priority Associated with 'Coronavirus' Collocates in COCA.

As well as identifying the lexical part-of-speech collocates of 'coronavirus', the COCA's interface has the pedagogically effective feature of what we prefer to call the corpus-driven topical priority associated with the collocates. As demonstrated in Fig. 9, the topical priority generated by the COCA and regarded as thematically relevant to the collocational pairs identified in Fig. 7 consists largely in specific topical domains: virus, acute, respiratory, disease, contact, coronaviruses, novel, and severe. Further, as presented in Fig. 8, the node term 'coronavirus' is defined within the topical scope of {virology}, which can be said to reveal the semantic nature of 'coronavirus' as [+viral].

Thus, bringing together the last two steps of part-of-speech-bound collocates and their topical priority may well improve the EFL students' understanding of 'coronavirus' as a concept; that is, beyond the term as a de-contextualized lexical item that is isolated from its significant collocates.

VI. Conclusion and Future Research

In conclusion, we are in a position to round off the pedagogical strategy proposed in the present study for enhancing EFL students' COCA-induced collocational usage of 'coronavirus'. The approach used towards the fulfilment of this goal has been presented as more corpus-driven than corpus-based. The COCA has been utilized for empirically validating the proposed strategy. Such a strategy can be said to have yielded three results. First, a node word can be semantically defined in relation to its potential collocates provided there should be (a) a lexically orientated part-of-speech framing of these collocates and (b) a genre-sensitive balanced set of data manipulated by corpus software. The present case in point was presented in the nominal form 'coronavirus' whose lexical collocates were statistically calculated and formally recognized as nouns, adjectives, verbs, and adverbs in the COCA.

Second, a two-stage investigation of the 'coronavirus' node-collocate relation has been undertaken in a pedagogically systematic fashion. The first stage was a demonstration of the COCA's interface and its main functions of sorting and limiting the searches for a particular term via grammatically annotated settings of POS tagging as well as other relevant functions of specifying frequency and MI collocation statistics. The second stage was provided as a pedagogical praxis with specific highlights: (i) setting the collocation default range ±4, (ii) constructing coronavirus frequency distribution over the COCA's genre-based sections (TV/Movies, Blog, Web-General, Spoken, Fiction, Magazine, Newspaper, and Academic), (iii) generating the top collocates of 'coronavirus' and their lexical parts of speech in the COCA, and bringing out the topical priority associated with 'coronavirus' collocates in the same corpus data.

Third, on a rather empirical level, the actual collocates of the node word 'coronavirus' have been generated from the COCA as nouns, e.g. infection, probability, syndrome, sar, virus, and ferret; adjectives, e.g. novel, respiratory, and new; verbs, e.g. cause, identify, confirm, and ferret; and adverbs, e.g. closely, newly, previously, meanwhile, and widely. Further, on the same empirical level, collocation-induced topical priority was derived from the COCA in thematic relevance to the above collocates of 'coronavirus'; and they consisted in the following topical domains: virus, acute, respiratory, disease, contact, coronaviruses, novel, and severe. In view of such an empirical finding, with recurrent COCA generation of 'virus', the node 'coronavirus' has (perhaps unsurprisingly) been demonstrated to fall in the topical scope of {virology}.

References

[1] H. Lien, "The effects of collocation instruction on the reading comprehension of Taiwan college students," Unpublished doctoral dissertation, Indiana University of Pennsylvania, Pennsylvania, 2003.

[2] M. J. McCarthy, "A new look at vocabulary in EFL," *Applied Linguistics*, vol. 5, no. 1, pp. 12-22, 1984.

[3] I. S. Nation, *Teaching and learning vocabulary*. Boston: Heinle & Heinle Publishers, 1990.

[4] I. S. Nation, *Learning vocabulary in another language* (3rd ed.). Cambridge: Cambridge University Press, 2002.

[5] J. R. Nattinger, and J. S. DeCarrico, *Lexical phrases and language teaching*. Oxford: Oxford University Press, 1992.

[6] J. R. Nattinger, and J. S. DeCarrico, *Lexical phrases and language teaching* (2nd ed.). Oxford: Oxford University Press, 1997.

[7] S. Shih, and H. Wang, "The Relationship Between EFL Learners' Depth of Vocabulary Knowledge and Oral Collocation Errors," In proceedings of The 23rd International Conference on English Teaching and Learning in the Republic of China, pp. 964-977. Taipei, Taiwan: Kaun Tang International Publishing Ltd., 2006.

[8] M. Davies, The Corpus of Contemporary American English (COCA), Available online at https://www.english-corpora.org/coca/,2008-

[9] J. Sinclair, *Corpus concordance collocation*. Oxford: Oxford University Press, 1991.

[10] J. Sinclair, *Trust the text: Language, corpus, and discourse*. London, New York: Routledge, 2004.

[11] B. Altenberg, and S. Granger, "The grammatical and lexical patterning of MAKE in native and non-native student writing," *Applied Linguistics*, vol. 22, pp. 173-195, 2001. https://doi.org/10.1093/applin/22.2.173.

[12] P. Durrant, and N. Schmitt, N., "To what extent do native and non-native writers make use of collocations?," *IRAL-International Review of Applied Linguistics in Language Teaching*, vol. 47, pp. 157-177, 2009. doi:10.1515/iral.2009.007.

[13] S. Granger, "Prefabricated patterns in advanced EFL writing: Collocations and formulae," In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, pp. 79-100. Oxford: Oxford University Press, 1998.

[14] P. Howarth, "The phraseology of learners' academic writing," In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications*, pp. 161-186. Oxford: Oxford University Press, 1998.

[15] B. Laufer, and T. Waldman, "Verb–noun collocations in second language writing: A corpus analysis of learners," *English Language Learning*, vol. 61, pp. 647–672, 2011.

[16] N. Nesselhauf, "The use of collocations by advanced learners of English and some implications for teaching," *Applied Linguistics*, vol. 24, pp. 223-242, 2003.

[17] T. M. H. Nguyen, and S. Webb, "Examining second language receptive knowledge of collocation and factors that affect learning," *Language Teaching Research*, pp. 1-23, 2016. doi:10.1177/1362168816639619.

[18] M. Paquot, and S. Granger, "Formulaic language in learner corpora," *Annual Review of Applied Linguistics*, vol. 32, pp. 130-149, 2012. doi:10.1017/S0267190512000098.

[19] J. Bahns, and M. Eldaw, "Should we teach EFL students collocations?," *System*, vol. 21, pp. 101-114, 1993.

[20] D. Biskup, "L1 influence on learners' rendering of English collocations: A Polish/German empirical study," In P. J. L. Arnaud, and H. Bejoint (Eds.), *Vocabulary and Applied Linguistics*, pp. 85-93. London: Macmillan, 1992.

[21] M. Farghal, and H. Obiedat, "Collocations: A neglected variable in EFL," *International Review of Applied Linguistics*, vol. 33, pp. 315-331, 1995.

[22] A. Hasselgren, "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary," *International Journal of Applied Linguistics*, vol. 4, pp. 237–258, 1994. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x.

[23] J. Li, & N. Schmitt, "The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach," In D. Wood (Ed.), Perspectives on Formulaic Language: Acquisition and Communication, pp. 22-46. New York: Continuum, 2010.

[24] L. Fang, Q. Ma, and J. Yan, "The effectiveness of corpus-based training on collocation use in L2 writing for Chinese senior secondary school students," *Journal of China Computer-Assisted Language Learning*, vol. 1, no. 1, pp. 80-109, 2021. https://doi.org/10.1515/jccall-2021-2004.

[25] H. C. M. Hu, "A semantic prosody analysis of three adjective synonymous pairs in COCA," *Journal of Language and Linguistic Studies*, vol. 11, no. 2, pp. 117-131, 2015.

[26] G. Kartal, and G. Yangineksi, "The effects of using corpus tools on EFL student teachers' learning and production of verb-noun collocations," *PASAA*, vol. 55, pp. 100-122, 2018.

[27] D. M. Mansour, "Using COCA to Foster Students' Use of English Collocations in Academic Writing," In proceedings of the 3rd International Conference on Higher Education Advances, HEAd'17 Universitat Politecnica de Valencia, Valencia, pp. 600-607, 2017. DOI: http://dx.doi.org/10.4995/HEAd17.2017.5301.

[28] I. N. Oktavianti, and J. Sarage, "Collocates of 'great' and 'good' in the corpus of contemporary American English and Indonesian EFL textbooks," *Studies in English Language and Education*, vol. 8, no. 2, pp. 457-478, 2021. https://doi.org/10.24815/siele.v8i2.18594.

[29] Y. Wu, "Discovering collocations via data-driven learning in L2 writing," *Language Learning & Technology*, vol. 25, no. 2, pp. 192-214, 2021.

[30] F. Boers, J. Eyckmans, H. Kappel, H. Stengers, & M. Demecheleer, "Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test," *Language Teaching Research*, vol. 10, pp. 245-261, 2006. https://doi.org/10.1191/1362168806lr195oa.

[31] D. A. Cruse, *Lexical semantics*. New York: Cambridge University Press, 1986.

[32] M. Hoey, *Patterns of lexis in text*. Oxford: Oxford University Press, 1991.

[33] G. Kennedy, "Collocations: Where grammar and vocabulary teaching meet," *Language Teaching Methodology for the Nineties*, RELC, Anthology Series 24, 1990.

[34] A. P. Cowie, "The treatment of collocations and idioms in learners' dictionaries," *Applied Linguistics*, vol. 2, no. 3, pp. 223-235, 1981.

[35] C. J. Gledhill, *Collocation in science writing*. Tubingen: Gunter Narr Verlag, 2000.

[36] D. A. Cruise, "Language, meaning and sense: Semantics. In N. E. (Ed.), *An Encyclopedia of Language*, pp. 139-172. New York: Routledge, 1990.

[37] D. Gablasova, V. Brezina, and T. McEnery, "Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence," *Language Learning, A Journal of Research in Language Studies*, vol. 67, no. 1, pp. 155-179, 2017. https://doi.org/10.1111/lang.12225

[38] J. Bahns, "Lexical collocations: A contrastive view," *ELT Journal*, vol. 47, no. 1, pp. 56-63, 1993.

[39] T. Fontenelle, "Lexical functions in dictionary entries. In A. P. Cowie, Phraseology: theory, analysis, and applications, pp. 189-207. Oxford: Oxford University Press, 1998.

[40] W. E. Louw, "Irony in the Text or Insincerity in the Writer?," In M. Baker (Ed.), *The Diagnostic Potential of Semantic Prosodies." Text and Technology: In Honour of John Sinclair*, pp. 157-176. Amsterdam: John Benjamins, 1993.

[41] J. Hill, "Revising priorities: From grammatical failure to collocational success," In M. Lewis, *Teaching Collocations: Further developments in the lexical approach*, pp. 47-69. Hove: Language Teaching Publications, 2000.

[42] S. Hunston, *Corpora in applied linguistics*. Cambridge: Cambridge University Press, 2002.

[43] K. Matsuno, "Processing collocations: Do native speakers and second language learners simultaneously access prefabricated patterns and each single word?," *Journal of the European Second Language Association*, vol. 1, no. 1, pp. 61–72, 2017. DOI: https://doi.org/10.22599/jesla.17.

[44] J. Hill**,** and M. Lewis (Eds.), *LTP dictionary of selected collocations*. Hove: Language Teaching Publications, 1997.

[45] M. A. K. Halliday, and R. Hasan (Eds.), *Cohesion in English*. Essex: Longman, 1976.

[46] C. J. Fillmore, P. Kay, and M. C. O'Connor, "Regularity and idiomaticity in grammatical constructions: The case of letalone," *Language*, vol. 64, pp. 501-538, 1988. https://doi.org/10.2307/414531.

[47] M. Stubbs, *Words and phrases*. Oxford: Blackwell, 2001.

[48] A. Wray, *Formulaic language and the lexicon*. Cambridge, England: Cambridge University Press, 2002.

[49] B. Laufer, "The influence of L2 on L1 collocational knowledge and on L1 lexical diversity in free written expression," In V. Cook (Ed.), *Effects of the Second Language on the First*, pp. 120-141. Clevedon: Cromwell Press Ltd, 2003.

[50] B. Erman, F. Forsberg Lundell, M. Lewis, "Formulaic language in advanced second language acquisition and use," In K. Hyltenstam (Ed.), *Advanced Proficiency and Exceptional Ability in Second Languages*, pp. 111-148. Boston: Walter de Gruyter, 2016.

[51] C. F. Meyer, "Corpus-based and corpus-driven approaches to linguistic analysis: One and the same?," In I. Taavitsainen, M. Kytö, C. Claridge, and J. Smith (Eds.), *Developments in English: Expanding Electronic Evidence*. Cambridge: Cambridge University Press, 2017.

[52] E. Tognini-Bonelli, *Corpus linguistics at Work*. Amsterdam: John Benjamins, 2001.

[53] R. Moon, *Fixed expressions and idioms in English*. Oxford: Oxford University Press, 1998.

[54] J. Yamashita**,** and N. Jiang, "L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations," *TESOL Quarterly*, vol. 44, no. 4, pp. 647-668, 2010. DOI: https://doi.org/10.5054/tq.2010.235998.

[55] N. Nesselhauf, *Collocations in a learner corpus*. Amsterdam: John Benjamins, 2005.

[56] T. Johns, "From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning," In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar (Cambridge Applied Linguistics*, pp. 293-313. Cambridge: Cambridge University Press, 1994.

[57] P. Pérez-Paredes, and G. Mark, *Beyond concordance lines: Corpora in language education.* John Benjamins Publishing Company, 2021.

[58] D. Crystal, *Language and the Internet* (2nd ed.). Cambridge: Cambridge University Press, 2006.

AUTHORS' PROFILE

Amir H.Y. Salama is currently Associate Professor of linguistics and English language in the Department of English, College of Social Science and Humanities in Al-Kharj, Prince Sattam Bin Abdulaziz University, Saudi Arabia. Also, he is a standing Professor of linguistics and English language in the Faculty of Al-Alsun (Languages), Kafr El-Sheikh University, Egypt. In 2011, Prof. Salama got his PhD in linguistics from the Department of Linguistics and English language at Lancaster University, UK. Since then, he has published at international journals like Discourse & Society, Critical Discourse Studies, Pragmatics and Society, Cogent Arts and Humanities, Semiotica, Corpora, and Translation Spaces. His research interests are systemic functional grammar, corpus linguistics, discourse analysis, pragmatics, cognitive semantics, translation studies, and semiotics.

ORCID: https://orcid.org/0000-0001-9320-558X

Waheed M. A. Altohami is currently Assistant Professor of English Language and Linguistics in the Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University (KSA). Also, he is a standing lecturer of English Language and Linguistics in the Department of Foreign Languages, Faculty of Education, Mansoura University, Egypt. His research interests include discourse analysis, cognitive semantics, corpus linguistics, and translation.

ORCID: https://orcid.org/0000-0001-8742-1366

A Computational Approach to Decode the Pragma-Stylistic Meanings in Narrative Discourse

Ayman Farid Khafaga[1], Iman El-Nabawi Abdel Wahed Shaalan[2]
College of Science & Humanities, Prince Sattam bin Abdulaziz University, Saudi Arabia[1, 2]
Faculty of Arts &Humanities, Suez Canal University, Egypt[1]
College of Education, Al-Azhar University, Cairo, Egypt[2]

***Abstract*—This paper presents a computer-based frequency distribution analysis to decode the pragma-stylistic meanings in one of the narrative discourse represented by Orwell's dystopian novel Animal Farm. The main objective of the paper is to explore the extent to which computer software contribute to the linguistic analysis of texts. The paper uses the variable of frequency distribution analysis (FDA) generated by concordance software to decode the pragmatic and stylistic significance beyond the mere linguistic expressions employed by the writer in the selected data. Some words were selected to undergo a frequency distribution analysis so as to highlight their pragmatic and linguistic weight which, in turn, helps arrive at a comprehensive understanding of the thematic message intended by the writer. The paper is grounded on one analytical strand: Frequency distribution analysis conducted by concordance. Results reveal that applying a frequency distribution analysis to the linguistic analysis of large data fictional texts serves to (i) identify the various types of discourse in these texts; (ii) create a thematic categorization that is based on the frequency distribution analysis of specific words in texts; and (iii) indicate that not only high frequency words are indicative in the production of particular pragmatic and stylistic meanings in discourse, but also low frequency words are highly indicative in this regard. These results accentuate a further general finding that computer software contribute significantly to the linguistic analysis of texts, particularly those pertaining to literature. The paper recommends further and intensive incorporation of computer and CALL (computer-assisted language learning) software in teaching and learning literary texts in EFL (English as a foreign language) settings.**

Keywords*—*Frequency distribution analysis; narrative discourse; pragma-stylistic meanings; thematic categorization

I. Introduction

For a long time, the use of computer-aided text analysis software has proven useful and contributive to the linguistic investigation of texts, particularly literary texts [1]. The reason why computer-aided text analysis (henceforth, CATA) tools are positively constructive in the analysis of literary genres lies in the fact that this type of texts always abound in large amount of data, i.e. a huge and vast number of lexis that it would be difficult to be analyzed without the help of computer [2], [3]. This paper, therefore, attempts to explore the pragma-stylistic meanings pertaining to George Orwell's Animal Farm, by using a frequency distribution analysis (FDA), which is considered one of the variables of CATA, to generate the occurrences of particular selected words, which, in turn, helps decode the various pragmatic and stylistic meanings pertaining to the text under investigation. The employment of CATA facilitates the whole process of readership and text reception on the part of readers, on the one hand, and helps clarify the different pragmatic and stylistic meanings encoded in the specific usage of each linguistic expression, on the other [4]. Linguistic expressions, in the context of this paper, include the different linguistic units, such as the word, the phrase, and the sentence. However, the analytical focus here will be on the level of the word. That is, some particular words are selected from the novel under investigation to undergo a frequency distribution analysis, by generating the total occurrences they have in the text, and the extent to which their frequency influences the general message of the text, the pragmatic meanings beyond the semantic proposition of the linguistic expressions, and the stylistic features adopted by the writer and textually reflected in the selected text.

A. Research Significance

The significance of this paper lies in its attempt to highlight the effective work of computer software in the linguistic analysis of large data texts, particularly fictional texts, by demonstrating the extent to which the pragmatic and stylistic meanings of some words can be revealed by the application of these computer software, such as the frequency distribution analysis adopted in this paper. This serves to emphasize the importance of applying and using the different computer software in the linguistic and textual analysis of texts. It is also anticipated to contribute to EFL (English as a foreign language) settings, computational linguistics, and corpus linguistics studies.

B. Research Questions

The current study attempts to answer the following research questions:

1) To what extent does a computer-based frequency analysis contribute to clarifying the pragma-stylistic meanings pertaining to particular words in the selected novel?

2) How does a frequency distribution analysis communicate specific pragmatic and stylistic meanings in the novel under investigation?

3) How do high frequency words mirror the thematic and discursive nature of the analyzed text?

4) Does the use and application of a frequency distribution analysis facilitate an intelligible perception and understanding of various types of texts?

C. Research Objectives

By applying a frequency distribution analysis to a number of selected words from the novel at hand, the study tries to achieve the following research objectives:

1) To explore the extent to which the use and application of computer software contribute significantly to the linguistic analysis of large data texts, particularly the fictional ones, as is the case with the novel under investigation.

2) To highlight the importance of incorporating the various computer software into the readership process of fictional texts.

3) To demonstrate the significance of applying computer software to the general understanding of the pragmatic, stylistic and thematic messages of literary texts.

The remainder of this article is structured as follows. Section II presents the literature review pertinent to the current study, as well as the related literature and the previous studies that addressed the same topic. Section III offers the methodology of the study, in which the selected data is described and the adopted analytical procedures are provided. Section IV demonstrates the analysis and results of the paper. Section V is dedicated to the discussion of the obtained results. Section VI is the conclusion, wherein some recommendations for future research are offered.

II. Literature Review

This section presents the review of literature as well as some previous studies that are relevant to the current study.

A. Computer-Aided Text Analysis (CATA)

The computer-aided text analysis offers a variety of analytical variables that can be employed in the analysis of the different types of texts [5], [6]. Various types of computer software are used in the linguistic analysis of texts, as they provide the analysts with different analytical options that help in deciphering the meanings beyond the linguistic expressions [7], [8]. According to [9], before the emergence and applications of modern technologies in teaching and learning, particularly in EFL settings, the analysis of large data texts such as the literary ones as well as the process of teaching them was very difficult. This is because the teacher and the student alike had to read the whole text in order to arrive at the meaning s/he targets, or what is called the thematic message of the text. The traditional way of reading literary texts, their reception on the part of students and their presentation on the part of teachers constituted an academic burden on both parties [10]. Furthermore, [11] emphasizes that the majority of results pertaining to the traditional analysis of literary texts are inaccurate. There is always some sort of mistake, specifically in terms of approaching the number of occurrences a given lexical item or a linguistic expression has in text. Traditional way of analyzing literary texts was a big problem, as it required much more effort and time than is the case if such analyses were conducted by means of CATA, with its different analytical options and software [12].

According to [13], the use of the computer and computational linguistics work makes it possible to process, access and examine large data for a diversity of purposes, and to investigate questions which could not plausibly be answered if the analysis was carried manually. He maintains that computer software provide various applications in automated indexing, classification, concordance, content analysis, thematic categorization and syntactic analysis. Furthermore, these computational software provide information on the company words keep in a corpus, and can also display a variety of senses of a lexical item type.

It is worth mentioning that CATA offers a number of analytical variables that contribute effectively to the process of analyzing large data texts in general, and literary genres in particular [14]. One of these variables is the frequency distribution analysis (FDA), which will analytically be applied in the current paper. An FDA serves to show the different frequencies and occurrences a given searched item has in a text [15]. One advantage of this variable is its ability to offer accurate, credible, authentic and concise results that rest beyond any supposed analytical vulnerability. According to [15], this high level of verification, credibility and authenticity attained by the application of FDA reflects the relevance of applying computer software to the analysis of the different types of texts.

Another analytical variable provided by CATA is what is called Key Word in Context (KWIC) [16]. This variable functions to offer the contextual environment in which the searched word occurs, which, in turn, serves to extend the pragmatic and stylistic purposes beyond the surface usage of this word. For [17], through KWIC, one can have a clear and credible picture of the preceding and subsequent words in company of the searched item. These preceding and subsequent contexts add more clarification to the pragmatic power of the linguistic expression, as well as to the way this single linguistic item contributes to the interpretation of the whole text.

According to [18], Content Analysis (CA) is another analytical variable that can be generated by CATA. This variable constitutes the process of analyzing the content produced, reproduced and maintained by any given lexical item. Significantly, by means of content analysis, one can delve into both the semantic compatibility of words; that is, the different semantic propositions these words communicate in the analyzed text and the pragmatic interpretation attributed to these words with their specific usages in texts.

A further variable offered by CATA in the context of this paper is related to the Thematic Categorization of discourse types [19]. This analytical option allows the analyst to categorize the various words with their frequencies with the themes they address in texts. Thematic distribution, therefore, is closely related to text clustering, as it identifies the different themes according to the number of frequencies words have. Crucially, this analytical variable operates in combination with KWIC; that is, it targets the thematic distribution of any searched word according to its contextual environment.

The last processing variable generated by CATA in this study is the variable of Lemmatization [20], a processing option which provides counts of lemmas (sets of grammatical words having the same stem and / or meaning and belonging to

the same major word classes, differing only in inflection and/or spelling. For [21], lemmatization serves to offer a classification of all the identical or related forms of a word under a common headword, which further functions to give a clear picture to the interpretative atmosphere of the analyzed text.

Significantly, all the aforementioned processing options can be generated by concordance, a computer software that allows researchers to access and process large data texts to produce the occurrences of particular tokens. According to Sinclair [22], the process of producing and accessing word indexes and concordances is the most obvious, conceivable and plausible application of the computer software in literary research. He emphasizes that such an automated approach to text analysis sets the analytical foundation for theoretical, empirical and analytical decisions on the various linguistic aspects of vocabulary expression, morphological, syntactic, and semantic dimensions of contained data, and the presentation of lexical and syntactic items and collocations of both high and low frequency tokens [23].

B. Previous Studies

Much previous research has been conducted on the application of CATA software into the linguistic investigation of texts. One study was presented by [24], who employed a computational approach that is based on a frequency distribution analysis to decode the extremist ideologies in the discourse of ISIS (Islamic State in Iraq and Syria). The study utilized the program of concordance to arrive at the total frequencies of specific words and collocations that help in the ultimate analysis of the discourse of ISIS. This study concluded that computer software are very effective in the analysis of all types of texts, as they offer accurate and credible results upon which discourse analysts can predicate their thematic and ideological investigation.

Another study conducted by [25] provided a computer-aided text analysis within courtroom discourse. This study employed CATA to explore the persuasive strategies used in the legal discourse of attorneys in the trial of Moussaoui that occupied the world and public opinion during the 1990s. The study clarified that CATA software are effective in deciphering the various persuasive tools employed by the two attorneys of the trial. The study further demonstrated the extent to which the number of occurrences of specific expressions within particular contexts in courtroom facilitates the process of persuasion on the part of the attorneys, either during their conversational turns with the allocated judge of the court or with testimonies of the witnesses. The study recommends the application of CATA software to the analysis of the different legal texts.

A third study by [26] investigated the ideological agency exercised by speakers over their recipients. This study uses concordance software in the analysis of the data to show the frequency distribution of the function words that indicate agency. The study also clarified that concordance proves useful in demonstrating the indicative occurrences out of the total frequency of each lexical token. This study concluded that the concept of ideological agency is better revealed via the application of computer software manifested in concordance.

A further relevant study was the one conducted by [27], in which they explored the extent to which CALL (computer-assisted language learning) software is effective in the EFL contexts. This study is entirely based on testing the effectiveness of using the two computer programs of Snagit$_{TM}$ and Screencast on acquiring the skill of reading. The study revealed that the application of the two computer software serves to improve the academic level of students, by fostering the linguistic skills pertinent to the acquisition of the skill of reading. The study also reported that such technological incorporation into EFL course functions to develop not only the linguistic competence of EFL students, but also their communicative skills. This study recommended the application of the different CALL software to the different EFL courses, as they facilitate the process of teaching and learning on the part of both teachers and students.

The previous studies so far have employed CATA software into the linguistic analysis of texts. Some of these studies focused on fictional texts, whereas other studies have presented discussions on legal texts and EFL settings. One observation concerning related literature is that it did not use CATA software within the scope of pragmatics and stylistics; that is, none of the previous studies has employed CATA software to explore the different pragmatic or stylistic purposes in discourse. This last point is the core concern of the current study, which constitutes the research gap attempted to be addressed in this article. The current study, therefore, tries to fulfill in part this research gap, by showing the effectiveness of using a frequency distribution analysis (FDA) as an indicator of discourse type and thematic categorization, as well as an analytical identifier of both the indicative and/or non-indicative occurrences in a corpus.

III. METHODOLOGY

This section presents the methodology of the study which constitutes data collection and description, the procedures adopted in the analysis of the selected data, and the rationale beyond the study.

A. Data: Collection and Description

The data in this paper comprises one literary text written by George Orwell: Animal Farm. A number of words from the novel were selected to undergo a frequency distribution analysis, which in turn clarified the high frequency words and the low frequency words and the extent to which both groups of words are indicative in reflecting the pragma-stylistic meanings pertaining to the selected novel. The selected words revolve around the discourse types of equality and inequality; the themes of oppression, rebellion and violence; and the point of view of the writer. Clarifying the way these concepts were perceived by means of the application of an FDA serves to mirror both the pragmatic and stylistic purposes targeted beyond their usage in the novel.

B. Research Procedures

Three procedural stages were adopted in this research. First, the use of the computer software adopted here has involved the preparation of the selected work by scanning and storing it electronically in order to be ready for computational analysis.

Second, the selected words were highlighted to undergo a frequency distribution analysis to generate their frequency of occurrences in the text under investigation. This stage was followed by content and thematic categorization analysis, wherein a connection has been made between the occurrences of each selected word and its significance to the pragmatics and stylistics of the novel as a whole. The third stage constitutes the interpretation and explanation of the results, as well as to relate the obtained results with the pragma-stylistic meanings communicated in the novel at hand.

C. Rationale of the Study

There are three reasons that constitute the rationale beyond the selection of Animal Farm: first, the novel has two different types of discourse: the discourse of equality and the discourse of inequality. Thus the selection of some words to undergo a frequency distribution analysis is relevant to identify the type of discourse. Second, the novel also abounds in themes that can further be decoded and categorized by the frequency analysis of a group of selected words. Third, the novel witnesses a number of words that are highly indicative in the production of the total interpretation of its incidents despite the fact that some of these words are very low in frequency.

IV. Analysis and Results

A. FDA as Indicator of Discourse Type

The application of an FDA serves to demonstrate the type of discourse stylistically communicated by the writer of the novel. The number of occurrences of some particular words mirrors the type of discourse in which these words occur and address. In Animal Farm, there are two types of discourse: the discourse of equality and the discourse of inequality. Consider the following table.

TABLE I. Frequency Distribution Analysis to Show Type of Discourse

Lexical item	Total Frequency	Indicative Occurrences
comrades	55	23
rebellion	29	2
wisdom	2	2
man	21	11
remains	1	1
miserable	5	3
slavery	3	1
free	7	1
leader	8	5
sacrifice	4	2
cruelty	2	1
criminal	1	1
traitor	3	2
laborious	3	1
tactics	2	2
comrade	41	13
equal	8	2
rich	3	1
brothers	1	1
percent	3	3
welfare	2	1
necessary	13	1
nonsense	1	1
nothing	24	1
agent	3	3
maneuver	2	1
get rid of	2	1
let us	12	2
true	7	1
enemy	12	12
friendship	1	1
friend	4	2
equality	2	2

Table I shows a number of words with the occurrences they have in the novel. The table demonstrates that there are some words pertinent to the discourse of equality, including equality, friendship, comrades, rebellion, wisdom, brother, and free. The associative meaning of closeness, brotherhood, cooperation and solidarity these words carry is an indication that they are related to the discourse of equality. It is also obvious from the table that there are some words among this group that have high frequency (e.g., rebellion, comrades), and other words that have low frequency (e.g., equality); however, the indication is that the word, regardless of its number of occurrences, may be very indicative in carrying the meaning of a specific type of discourse. In the same vein, Table I displays some words that can be perceived as indicators of an inequality discourse. Words such as man, miserable, cruelty, traitor, criminal, enemy, and remains also communicate the connotative meanings of oppression, domination and inequality. Again, the low frequency words of this group accentuate the fact that the words may be very low in frequency (e.g., slavery, laborious) but highly indicative in carrying the meaning of the discourse targeted by the writer.

B. FDA as Indicator of Thematic Categorization

The FDA can also be an indicator that clarifies the categorization of themes in texts. In the novel under investigation, one can identify a number of themes, such as violence, rebellion and oppression. These themes can be decoded by means of a frequency analysis that classifies the different words addressing a particular theme. Consider the following table.

TABLE II. FREQUENCY DISTRIBUTION ANALYSIS OF THE WORDS CONVEYING THE CONNOTATIVE MEANINGS OF 'VIOLENCE'

The Word	Frequency
kill	4
slaughtered	2
attack	10
executed	2
blood	6
shot	4
destruction	1
confessed	6
killed	5
destroyed	5
crush	1
suicide	1
tortured	1
torn	2
tore	1
executions	2
slaughter	1
bloodshed	1

Table II displays the words that encode the associative meaning of the lexical item 'violence'. One observation is that all words included in the above table are ideologically-loaded. That is, they are carriers of specific meanings pertaining to specific theme, which is violence. The words in the table direct the readers to one meaning, that is, there is violence among discourse participants. All words connote the violent meanings of torture and suffering.

The theme of 'rebellion' can also be detected by the FDA as is shown in the following table.

TABLE III. FREQUENCY DISTRIBUTION ANALYSIS OF THE WORDS CARRYING CONCEPTUAL AND ASSOCIATIVE MEANINGS OF 'REBELLION'

The Word	Frequency
rebellion	29
remove	1
expel	2
expulsion	8
uprising	1
striking	1
quarrel	2
quarrelling	1
dismissed	1
disobey	1
disobedience	1
rebelliousness	1
revolutionary	1
get rid of	2

Table III indicates that Orwell enriches the text with words indicating the meaning of rebellion, which is one of the main themes presented in Animal Farm. All lexis in the above table communicate the meaning of rebellion, both literally (e.g., rebellion, uprising, rebelliousness, revolutionary) and associatively (e.g., remove, get rid of, disobedience, dismissed, expel, striking). Further, it is not only high frequency words that communicate the theme of rebellion, but also low frequency words are carriers of the same theme.

A further theme presented in the novel can be also identified via the FDA and the number of occurrences of specific words. It is the theme of 'oppression' as is displayed in the following table.

TABLE IV. FREQUENCY DISTRIBUTION ANALYSIS OF THE WORDS CARRYING CONCEPTUAL AND ASSOCIATIVE MEANINGS OF 'OPPRESSION'

The Word	Frequency
destruction	1
slavery	3
miserable	5
cruelty	2
traitor	3
laborious	3
criminal	1
execution	1
kill	4
destroy	1
execute	2
destroyer	1
dominate	3

Table IV clarifies the various words carrying the literal and connotative meanings of oppression. The semantic potentials of the list of words in the above table refer to the theme of oppression and domination. All words in this list carry meanings that indicate suffering, submission and dominance; such meanings are much more pertinent to the theme of oppression than to any other themes in the discourse of the novel.

C. FDA as Identifier of Indicative / Non-indicative Occurrences

This part of the analysis sheds light on one important idea beyond FDA adopted in this paper, that is, the significance of both high frequency and low frequency words in communicating the various pragmatic and stylistic purposes in discourse. Unlike the general assumption that only high frequency words are significant in delineating the various discursive aspects of texts, the paper shows that low frequency words have the same significance in producing and maintaining specific discourse meanings and themes. Consider the following Tables V and VI.

TABLE V. INDICATIVE HIGH FREQUENCY WORDS IN ANIMAL FARM

The Word	Frequency
comrades	55
enemy	12
comrade	41
work	72
rebellion	29
man	21
all	174
no	102

TABLE VI. INDICATIVE LOW FREQUENCY WORDS IN ANIMAL FARM

Word	Frequency
equal	8
friends	5
blood	6
equality	2
revolutionary	1
uprising	1
expulsion	8
remove	1
friendship	1

Tables V and VI demonstrate a number of words with high frequency occurrences (Table V), and other words with low frequency occurrences (Table VI). In both cases, the words contribute to the general interpretation of the text under investigation. For example, the words work, comrades and rebellion show high frequencies of 72, 55, and 29, respectively. The indication here is that the high frequency of occurrences pertaining to these words communicates various discourse meanings, such as equality, cooperation and friendship. These meanings, in turn, are indicators of the type of discourse as well as the theme intended to be conveyed pragmatically and/or stylistically by the writer. Likewise, the very low frequency of occurrences of words, such as equal, uprising, blood, remove and expulsion (Table VI) does not mean that these words are insignificant and, thus, do not contribute to the discourse meanings of the novel. Contradictorily, these words, however, low in frequency, are highly contributive to the identification of the thematic message of the novel at hand.

D. FDA as Lemmas Generator

A frequency distribution analysis can further be perceived as a generator of lemmatization. These lemmas are also indicators of the pragma-stylistic meanings targeted in the novel, as they refer to the various discursive and thematic purposes beyond the text in which they occur. Consider the following Tables VII and VIII.

TABLE VII. LEMMATIZATION IN ANIMAL FARM

Destroy lemma		Execute lemma		Kill lemma		confess lemma	
word	**F**	**word**	**F**	**word**	**F**	**word**	**F**
destroy	1	executed	2	kill	4	confess	2
Destroyed	5	execution	1	kills	1	confessed	6
Destroyer	1	execution-s	2	killed	5	confessing	1
destroying	1			killing	1	confession	1
destruction	1					confessions	1

(F) means frequency

TABLE VIII. LEMMATIZATION IN ANIMAL FARM

equality lemma		friend lemma		comrade lemma		rebellion lemma	
word	**F**	**word**	**F**	**word**	**F**	**word**	**F**
equal	8	friend	4	comrade	34	rebellion	29
equally	3	friends	2	comrades	46	rebellions	1
equality	2	friendly	3	comrade-ship	1	rebelling	1
		friend-ship	1			rebelliou-sness	1

(F) means frequency

The two tables display the lemmas of various indicative words in the discourse of the novel. These lemmas are indicators of the stylistic way of writing adopted by the writer. The selection and use of these words, together with their different lemmas is dexterously employed to direct the cognitive background of the reader towards specific meanings that serve to arrive at the pragmatic interpretation of texts. Lemmas are very indicative in indentifying the type of discourse as well as in classifying the different themes in texts.

V. DISCUSSION

The above analysis demonstrates the effectiveness of using and applying CATA represented by FDA to the linguistic analysis of texts. It is analytically clarified that the use of modern technology in the linguistic and textual analysis of texts, particularly literary ones contributes effectively to the interpretation of these texts. In light of this paper, the application of FDA proves useful and contributive in creating and deleneating the general interpretative atmosphere of the novel under investigation. This variable of CATA facilitates the process of decoding the various pragmatic and stylistic meanings pertaining to the text at hand. Themes such equality, violence, oppression and rebellion have computationally been decoded by means of FDA. This correlates with a number of previous studies, such as [28], [29] and [25], which emphasize the importance of applying modern technologies to the textual and linguistic analysis of texts. Crucially, computer software, when used in corpus linguistic, function to facilitate the process of linguistic analysis, as they help make texts more manageable analytically (Research question No. A: to what extent does a computer-based frequency analysis contribute to clarifying the pragma-stylistic meanings pertaining to particular words in the selected novel?).

Pragmatically, the application of CATA software in general and FDA in particular proves useful in deciphering the various pragmatic meanings pertaining to the discourse of Animal Farm. Pragmatic meanings are meant to the implied or the invisible meanings. This is conducted by using FDA as an indicator to mirror the intended meanings targeted by the writer. For example, FDA has shown certain pragmatic meanings relevant to the idea of totalitarianisn, oppression and submission, which represent the main meanings intended beyond the surface meanings of the linguistic expressions in the novel [30]. Deciphering the intended meanings of the work under investigation by means of FDA also serves to reflect the intention of the writer, which is further meant by the ideological point of view (Research question No. B: How does a frequency distribution analysis communicate specific pragmatic and stylistic meanings in the novel under investigation?).

Stylistically, the use of FDA demonstrates the way the writer employes certain lexis that communicate specific meanings beyond the style of his writing. The analysis shows two types of discourse: the discourse of equality and the discourse of inequality. Each type of discourse is featured by a number of stylistic devices manifested in the clever use of specific words that direct the interpretative wheel of the text towards the targeted type of discourse. Consequently, one conclusion can be drawn here, that is, CATA software can be used to determine the type of discourse in texts. This finding goes in conformity with some previous studies, such as [20], [23] and [31], who clarified that the application of CATA to the linguistic study of literary texts serves to identify the nature of discursivity in literary genres (Research question No. B: How does a frequency distribution analysis communicate specific pragmatic and stylistic meanings in the novel under investigation?).

Thematically, the application of FDA helps classify the different themes addressed in literary texts. Such thematic categorization is highly required, particularly in literary genres, for the very nature these texts have concerning the large data they contain. The thematic classification can also facilitate the process of teaching and learning literary courses. This reconciles with [32], who emphasizes the significant contribution computer software can present to EFL settings. The application of these software can save time and effort on the part of both teachers and students. It also tunes with [33], who argue that CATA software prove contributive to the acquisition of the different language skills, particulaly, reading, writing, as well the acquistion of vocabulary. The use of computer software serves to improve students' performance and competency (Research question No. C: How do high frequency words mirror the thematic and discursive nature of the analyzed text?).

Crucially, the huge technological development necessitates the integration of computer softwar not only in EFL settings, or in the linguistic study of literary texts, but also in the linguistic investigation of further types of texts, such as legal and religious texts. The application of the various CATA software produce credible, authentic and concise results [34]. Using computer software also opens new analytical and pedagogical insights that would be difficult to be identified if the analysis is conducted without the help of computer. This last point was accentuated by [35], who shed light on the different theoretical, analytical and pedagogical horizons computer software offer for researchers in the different fields of the academic and scientific research (Research question No. D: Does the use and application of a frequency distribution analysis facilitate an intelligible perception and understanding of various types of texts?).

VI. Conclusion

This paper presented a computer-based frequency analysis to decode the pragma-stylistic meanings in Orwell's Animal Farm. The paper demonstrated the significance of using and applying computer software in general and FDA in particular to the linguistic study of literary texts. These software function to save both time and effort, as well as provide results that are more credible, accurate and concise than those arrived at by traditional ways of linguistic analysis (i.e. without the work of computer). The analysis of the current paper clarified that FDA proves useful in (i) identifying the types of discourse in the novel at hand; (ii) categorizing the various themes in discourse; and (iii) highlighting the indicative and non-indicative occurrences that communicate different pragmatic and stylistic purposes in the novel under investigation. These three analytically-evidenced findings are computationally enabled by the application of FDA. The paper emphasizes the findings revealed by previous studies, by highlighting the significance and necessity of using the various computer software in the linguistic analysis of texts, particularly large data fictional texts. This is because these software facilitate the whole process of analysis, open new analytical horizons in the field, improve textual and contextual intelligibility pertaining to texts, provide fast, credible and concise results, and mirror the pragmatic and stylistic meanings communicated by writers.

Finally, for future research, this paper recommends further applications of the different computer software to the analysis of texts other than the literary ones. For example, to investigate the effectiveness of GBL (Game-Based Learning) on the performance of EFL (English as a Foreign Language) majors concerning vocabulary acquisition, or investigating the impact of CAT (Computer Assisted Translation) Trados Studio software on the teaching and learning translation. These recommended studies might reveal similar and/or different results than those approached in the current paper. Crucially, integrating computer software in the EFL settings contributes significantly to the teaching methods on the part of instructors, and to the learning outcomes on the part of students.

Acknowledgment

The authors take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship for the technical support it has unstintingly provided towards the fulfillment of the current study.

References

[1] G. Stockwell, Computer-assisted language learning: Diversity in research and practice. Cambridge: Cambridge University Press, 2018.

[2] A. F. Khafaga, "The perception of blackboard collaborate-based instruction by EFL majors/teachers amid COVID-19: A case study of Saudi universities". Journal of Language and Linguistic Studies, vol. 17, no. 2, pp. 1160-1173, 2021.

[3] D. Wiechmann, and S. Fuhs, "Concordancing software". Corpus Linguistics and Linguistic Theory, vol. 2, no. 2, pp. 107-127, 2006.

[4] D. Krieger, "Corpus linguistics: What it is and how it can be applied to teaching". The Internet TESL Journal, vol. IX, no. 3, pp. 123-141, 2003.

[5] K. Romeo, "A web-based listening methodology for studying relative clause acquisition". Computer Assisted Language Learning, vol. 21, no. 1, pp. 51-66, 2008.

[6] J. Flowerdew, "Concordancing as a tool in course design". System, vol. 21, no. 2, pp. 231-244, 1993.

[7] F. Yavus, "The use of concordancing programs in ELT". Procedia-Social and Behavioral Sciences, vol. 116, pp. 2312-2315, 2014.

[8] I. Pollach, "Taming textual data: The contribution of corpus linguistics to computer-aided text analysis". Organizational Research Methods, vol. 15, no. 2, 263–287, 2012.

[9] H. Bergqvist, "Swedish modal particles as markers of engagement: Evidence from distribution and frequency". Folia Linguistica, vol. 54, no. 2, pp. 469-496, 2020.

[10] A. F. Khafaga, and I. Shaalan, "Mobile learning perception in the context of COVID-19: An empirical studyof Saudi EFL majors". Asian EFL Journal, vol. 28, no. 1.3, pp.336-356, 2021.

[11] J. Reddington, F. Murtagh, and C. Douglas, "Computational properties of fiction writing and collaborative work". International Symposium on Intelligent Data Analysis, pp. 1-13, 2013.

[12] Q. Ma, "From monitoring users to controlling user actions: A new perspective on the user-centred approach to CALL". Computer Assisted Language Learning, vol. 20, no. 4, pp. 297-321, 2007.

[13] S. Hockey, A guide to computer applications in the humanities. London: The Johns Hopkins University Press, 1980.

[14] G. Kennedy, An introduction to corpus linguistics. London & New York: Longman, 1998.

[15] A. Barger, and K. Byrd, "Motivation and computer-based instructional design". Journal of Cross-Disciplinary Perspectives in Education, vol. 4, no. 1, pp. 1-9, 2011.

[16] K. Beatty, Teaching and researching computer-assisted language learning. Harlow: Longman Pearson, 2010.

[17] A. F. Khafaga, and I. Shaalan, "Using concordance to decode the ideological weight of lexis in learning narrative literature: A computational approach". International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, pp. 246-252 , 2020.

[18] A. F. Khafaga, and I. Shaalan, "Pronouns and modality as ideology carriers in George Orwell's Animal Farm: A computer-aided critical discourse analysis". TESOL International Journal, vol. 16, no. 4.2, pp. 78-102, 2021.

[19] R. Dzekoe, "Computer-based multimodal composing activities, self-revision, and L2 acquisition through writing". Language Learning & Technology, vol. 21, no. 2, pp.73-95, 2017.

[20] A. F. Khafaga, Strategies of political persuasion in literary genres: A computational approach to critical discourse analysis. Germany: LAMBERT Publication, 2017.

[21] A. Thabet, "Applied computational linguistics: An approach to analysis and evaluation of EFL materials". Damietta Faculty of Education Journal, vol. 1, no. 13, pp. 7-39, 1990.

[22] J. Sinclair, Corpus, concordance collocation. Oxford: Oxford University Press, 1991.

[23] A. F. Khafaga, "Exploring ideologies of function words in George Orwell's Animal Farm". Pertanika Journal of Social Sciences and Humanities, vol. 29, no. 3, pp. 2089 -211, 2021.

[24] A. F. Khafaga, "A computational approach to explore the extremist ideologies of Daesh discourse". International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 193-199, 2020.

[25] A. F. Khafaga, and B. Aldossari, The language of persuasion in courtroom discourse: A computer-aided text analysis". International Journal of Advanced Computer Science and Application, vol. 11, no. 7, pp. 332-340, 2021.

[26] A. F. Khafaga, and M. Aldawsari, "Ideological agency in Edward Bond's Lear". Applied Linguistics Research Journal, vol. 5, no. 2, pp. 11-23, 2021.

[27] A. F. Khafaga, and A. Alghawli, "The impact of CALL software on the performance of EFL students on the Saudi university context". International Journal of Advanced Computer Science and Application, vol. 12, no. 7, pp. 304-312, 2021.

[28] C. A. Chapelle, and J. Jamieson, Tips for teaching with CALL: Practical approaches to computer-assisted language learning. White Plains, NY: Pearson Education, 2008.

[29] L. Pedro, C. Barbosa, and C. Santos, "A critical review of mobile learning integration in formal educational contexts". International Journal of Educational Technology in Higher Education, vol. 15, no. 1, 2018.

[30] W. F. Bolton, The Language of 1984. Orwell's English and Ours. Basil Blackwell Publisher Limited, 1984.

[31] J. Flowerdew, "Concordancing as a tool in course design". System, vol. 21, no. 2, pp. 231-244, 1993.

[32] C. Pim, "Emerging technologies, emerging minds: Digital innovations within the primary sector," in: G. Motteram (Ed.), Innovations in learning technologies for English language teaching, London: British Council, pp. 17-42, 2013.

[33] J. Jarvis, and I. Pastuszka, "Electronic literacy reading skills and the challenges for English for academic purposes". CALL-EJ Online, vol. 10, no. 1, 2008.

[34] R. S. Pinner, "Teachers' attitudes to and motivations for using CALL in and around the language classroom". Procedia-Social and Behavioral Sciences, vol. 34, pp. 188-192, 2012.

[35] A. Oskoz, and I. Elola, "Digital stories: Overview". CALICO Journal, vol. 32, no. 2, pp. 155-173, 2016.

AUTHORS' PROFILE

Ayman Khafaga is an Associate Professor of Linguistics at the department of English, College of Science &Humanities, Prince Sattam Bin Abdulaziz University, Saudi Arabia. He is also an Associate Professor of Linguistics, Faculty of Arts & Humanities, Suez Canal University, Egypt. His research interests include computational linguistics, discourse studies, semantics, pragmatics and stylistics. ORCID: https://orcid.org/0000-0002-9819-2973.

Iman El-Nabawi Shaalan is an Associate Professor of Applied Linguistics at the College of Science and Humanities, Department of English , Prince Sattam bin Abdul-Aziz University, Saudi Arabia. She is also an Assistant Professor of TEFL at the College of Education, Al-Azhar University, Cairo, Egypt. Her research interests include language teaching methodologies, curriculum development and teaching methods, second and foreign language teaching and learning, and translation. ORCID: https://orcid.org/0000-0002-5411-7613.

An Evaluation of the Automatic Detection of Hate Speech in Social Media Networks

The Case of Arabic Posts on Facebook Regarding France's Muhammad Cartoon Controversy

Abdulfattah Omar[1]
Department of English, College of Science & Humanities
Prince Sattam Bin Abdulaziz University, KSA
Faculty of Arts, Port Said University, Egypt

Mohamed Elarabawy Hashem[2]
Department of English, College of Science and Arts in Tabarjal, Jouf University, KSA
Al-Azhar University, Cairo, Egypt

Abstract*—Numerous approaches have been developed over recent years to detect hate speech on social media networks. Nevertheless, a great deal of what is generally recognized as hate speech cannot yet be detected. There remain many challenges to assuring the effectiveness and reliability of automatic detection systems in different languages, including Arabic. Social media platforms and networks such as Facebook continue to encounter difficulties regarding the automatic detection of hate speech in Arabic content. Given the importance of developing reliable artificial intelligence and automatic detection systems that can reduce the problems and crimes associated with the spread of hate speech on social media platforms, this study is concerned with evaluating the performance of the automatic detection and tracking of hate speech in Arabic content on Facebook. As an example, the study evaluates the period in October 2020 that came to be known as France's cartoon controversy. Two different corpora were designed. The first corpus comprised 347 posts deleted by Facebook, now known as Meta. The second corpus was composed of 1,856 posts that were randomly selected using the hashtag إلا رسول الله (except the Prophet of Allah). The results indicate that there is a considerable amount of hate speech taken from or influenced by the Islamic religious discourse, but that automatic detection systems are unable to address the peculiar linguistic features of Arabic. There is also a lack of clarity in defining what constitutes "hate speech". The study suggests that social media networks, including Facebook, need to adopt more reliable automatic detection systems that consider the linguistic properties of Arabic. Political thinkers and religious scholars should be involved in defining what constitutes hate speech in Arabic.

Keywords*—Artificial intelligence; automatic detection; Facebook; hate speech; Islamic discourse; social media networks

I. Introduction

In recent years, the spread of social media networks and platforms has resulted in the emergence of different forms of hate speech, which have negative impacts on the stability of societies [1]. Millions of users around the world today use these social media networks and platforms to spread hate against specific groups and individuals [2, 3]. It is clear that hate speech has a central role in various discussions, including those on immigration, politics, sports, religion, and even diseases [4-6]. Hate speech has also been associated with crime, racial hatred, and violence [7, 8]. In the face of the increasing threats posed by hate speech to the lives of individuals and societies, social media networks have adopted a range of automatic detection systems with capabilities in different languages, especially Indo-European languages [9]. For his part, Mark Zuckerberg, the Chief Executive of Facebook, expressed his commitment to addressing the issue of hate speech on the platform. In a speech made at the ceremony for the newly established Axel Springer Award in Berlin on 25 February, 2016, Zuckerberg stressed that "hate speech has no place on Facebook and in our community". In a recent report, Facebook announced that the company removed 22.3 million pieces of content containing hate speech, down from 31.5 million in the second quarter of 2021, as shown in Fig. 1.

However, a report by the Wall Street Journal in 2021 highlighted that Facebook removed posts that generated just 2% of the hate speech viewed on the platform and that violated its rules [10]. In the face of these contradictory statistics, many users, groups, and organizations have questioned Facebook's figures and thus the reliability of automatic detection and the artificial intelligence systems adopted by Facebook for detecting and tracking hate speech in its content. Many users have criticized the lack of effectiveness of the company's procedures for curbing hate speech on the platform, for instance allowing ISIS members and supporters to use it. In contrast, others have described the company as taking a Big Brother approach in dictating what can and cannot be said [11].

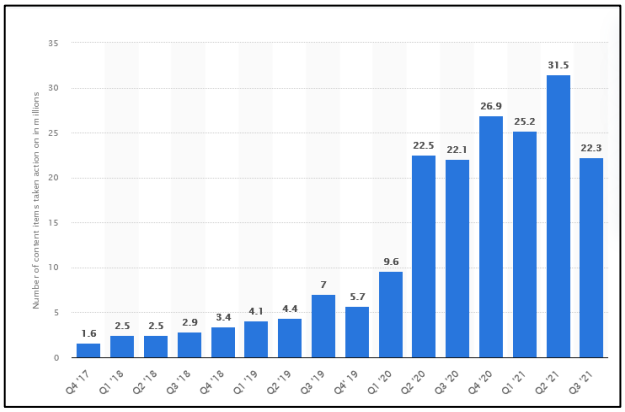


Fig. 1. Global Number of Hate Speech-containing Content removed by Facebook from 4th Quarter 2017 to 3rd Quarter 2021.

To illustrate the issue, this study evaluates the automatic detection of hate speech on Facebook in October 2020 during what came to be known as France's cartoon controversy. In October 2020, statements made by the French President Emmanuel Macron concerning Islam and the Prophet Muhammad led to many protests in the Arab and Muslim world. In these statements, Macron declared that his country would not stop publication of offensive cartoons of the Prophet, referring to them as freedom of expression. Macron's statements were warmly received by many activists, who described them as an assertion of France's "freedom to speak, to write, to think, to draw". Millions of Facebook users supported Macron's case, depicting Muslims as terrorists, especially after the brutal murder of a French teacher beheaded for showing his students cartoons of the Prophet Mohammed [12]. In turn, many commentators depicted Macron's statements as hate speech and a call for violence [13]. Furthermore, several hashtags trended in different Arab and Muslim countries through which activists described the statements of the French President as an insult to the Prophet of Islam and Muslims around the world. These hashtags included "except the Prophet of Allah", "boycott French products", "our prophet is a red line", "Macron offends the Prophet", and "stop insulting our Prophet". For its part, Facebook removed thousands of posts that were defined by the company as hate speech. In light of the above, this study seeks to evaluate the performance of artificial intelligence and automatic detection systems adopted by Facebook to understand how well they work and the extent to which they achieve their goals.

The remainder of this article is organized as follows. Section II provides a brief survey of automatic detection systems and approaches. Section III describes the methods and procedures. Section IV reports the results of the study. Section V is an interpretation of the results. Section VI concludes.

II. Related Work

Recent years have seen increasing interest in "hate speech" in research studies. The phenomenon has been extensively studied in various disciplines, including discourse studies, social media research, sociology, and recently artificial intelligence, data mining, and information studies. This can be attributed to the increasing rates of crimes associated with hate speech on social media networks and platforms. Although the concept of "hate speech" was evident in different societies before the emergence of social media networks and platforms, the concept has recently been linked to social media [14]. Despite the usefulness and reliability of these networks and platforms for bringing people closer to each other, they have unfortunately also helped to disseminate user-generated content that gives rise to hate speech on heated political and religious topics [15, 16].

In the face of this issue, researchers have sought to develop automatic detection systems and algorithms with the capability of identifying hate speech in content so that such posts can be removed [1, 17]. Studies in this tradition are usually multidisciplinary. That is, they are based on different disciplines, including artificial intelligence, data mining, natural language processing, and computational linguistics [18, 19]. The underlying principle is that algorithms should be trained to identify linguistic content and detect forms of hate speech through artificial intelligence and data mining tools [20, 21]. In this regard, linguistics research has always been central to the development of automatic detection systems. Capozzi et al. [22] argue that hate speech can be deployed through various morphological structures and lexical choices with a myriad of nuances geared to the context of situation. In some languages, dictionaries of terms used in hate speech have been compiled.

As noted by Cobbe [23], artificial intelligence systems can usefully be employed to control and monitor hate speech on social platforms. Fortuna and Nunes [24] similarly argue that automatic detection methods are effective mapping tools for tracking the diffusion of hate speech on a large scale across regions. Nonetheless, the detection of hate speech can be challenging for machines, let alone humans, due to the complexity of determining lexical referentiality [25]. Natural language processing designers have developed operational frameworks focusing on representative features and based on semantic classifications [26], but these always have to be linked to the context for the meaning of the lexis to be effectively attributed to the notion of hate speech [27].

The literature indicates that much automatic detection research has focused on social media networks and platforms, including Facebook and Twitter. Since these networks exhibit different forms of hate speech, they provide good opportunities for researchers to test their models in different languages, including English, Spanish, Italian, and Chinese [28]. For instance, Poletto et al. [29] used the Twitter platform for data collection to detect hate speech communicated by Italian users on social media with regard to immigrants. Similarly, Vigna et al. [30] examined the hateful content of speech presented on Facebook.

Although there is extensive literature on the automatic detection of hate speech in different languages, including English and Chinese, very little has been done in Arabic due to the linguistic differences between Arabic and Western languages. However, the considerable spread of hate speech and abusive language on social media in recent years has led to pressure on the industry and researchers to find workable and reliable solutions for hate speech problems in the Arab world.

According to Bahaa-eddin [31], the rise in hate speech on social media in Arab countries can be described as a "tsunami" that has grave consequences for the stability of Arab societies. He suggests that the unprecedented growth in hate speech in recent years can be ascribed to the intermittent, but ongoing turmoil in the region, such as the Iraqi invasion of Kuwait, the 9/11 attacks that left Arabs with diverse views, the war on Iraq, the Israeli–Palestinian conflict, the clashes between Shias and Sunnis, and very recently the Arab Spring with all its repercussions. All these events and more have had a significant effect on the temper of the Arab public. Within this environment, social media platforms allow domains in which people can comment and use insulting and offensive language in their interactions.

In this regard, there have been various attempts in recent years to develop automatic detection systems to address hate speech in Arabic. Al-Hassan and Al-Dossari [32], for instance,

used deep learning within artificial neural networks to build a model that mimics layers of neurons to identify patterns in the text. Likewise, Watanabe et al. [33] proposed the use of n-gram features for detecting hate speech on Twitter. In addition to these efforts, the study of hate speech in Arabic content on social media platforms still accelerates in many respects.

III. Methods, Data and Procedures

This study is based on two different corpora built from Facebook posts covering France's cartoon controversy in October 2020. The first corpus is composed of 1,347 posts deleted by Facebook, now known as Meta. The second corpus comprises 1,856 posts that were randomly selected using the hashtag إلا رسول الله (except the Prophet of Allah). Data were collected from October 18 through November 5, 2020. The study is limited to posts in Arabic.

The deleted posts from Facebook included terms that were described as of a threatening nature, as shown in Table I.

In the second corpus (based on the hashtag إلا رسول الله [except the Prophet of Allah]), posts were clustered using vector space clustering methods. The posts were classified into four main groups (clusters). The most distinctive lexical features of Cluster 1 included words such as coexistence, tolerance, understanding, values, peace, and mercy. The second cluster included words such as "terrorists", "murderers", "bloody", and "beasts". The third cluster included words such as "pigs", "Jews", "Christians", and "enemies". Finally, the last cluster included almost all the words in the third cluster and encompassing different writing styles.

TABLE I. List of Hate Speech Deleted by Facebook

Arabic Terms	English Translation
انزل غضبك ومقتك	may you pour out your wrath and hatred
يا حثالة	O scumbags!
يا كلاب	O dogs!
الكلب ماكرون	Macron, the dog
انزل غضبك و سخطك	May you pour out your wrath and hatred
انتقم منهم اشد انتقام	take revenge on them
المتصهينين	The Zionists
طبع الله علي قلوبهم	May Allah close off their hearts
لن يتغير اليهود	the Jews will never change
الخاسرين	the losers
قطع الله ألسنتكم	may Allah silence you
ابن الكلب	son of a dog
مواجهة التوسع الصهيوني و الإيديولوجية المسيحية	confronting Zionist expansion and Christian ideology
أزل دولتهم	Remove their country
سلط داء ليس له دواء	may you grant them a disease which has no cure
احفظ لنا ماكرون وابن اليهودية في ثلاجة الموتى	keep Macron, the son of Judaism, in the mortuary
ياعـبـيـد الـبقـر	O worshippers of cows
شل الإله لك اليمين	may Allah paralyze you
شلت أيديهم	may their hands be paralyzed
اللهم عليك بالظالمين	may you annihilate the wrongdoers
اللهم عليك بالكفر وأهله	may you wipe the unbelievers out
النازية	Nazism
شتت شملهم	dissolve their unity
فرق جمعهم	divide their gatherings

For the purposes of the study, Facebook's Policy Rationale developed for the definition of hate speech is adopted, as shown in Fig. 2.

> We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence.

Fig. 2. Facebook's Policy of Hate Speech.

IV. Results

As mentioned above, the posts in the second corpus were clustered into four distinct classes. To identify the thematic features of each group, a centroid-based lexical analysis was carried out. Based on Facebook's policies and definition of hate speech, Clusters 2, 3, and 4 are classified as hate speech and harmful content. Posts in these clusters constitute around 67% of the overall posts in the corpus, as shown in Table II.

It was clear that many users employed undefined writing systems to deceive Facebook's artificial intelligence algorithms. Arabic has a unique writing system, which is completely different from Western languages. In the Arabic orthographic system, dotting is a special characteristic that is used to address the problem of ambiguities in Arabic consonants [34]. According to Maroun [35], thirteen of the 28 Arabic letters include dots, which can be placed above or below letters. Some of these letters have one dot (e.g., ب /b/), while others have two (e.g., ي /j/) or three (e.g., /ʃ/ش). Sometimes, just one dot can distinguish between two or more words (e.g., جديد حديد /ħadi:d/,/ʤadi:d/ iron, new). Interestingly, Classical Arabic was used without dotting. According to Al-Azami [36], only context was used to identify the consonants, as shown in Fig. 3.

TABLE II. Classification of the Four Lexical Clusters

Cluster	*Number of posts*	*Percentage*
Cluster 1	618	33%
Cluster 2	137	7%
Cluster 3	837	46%
Cluster 4	264	14%

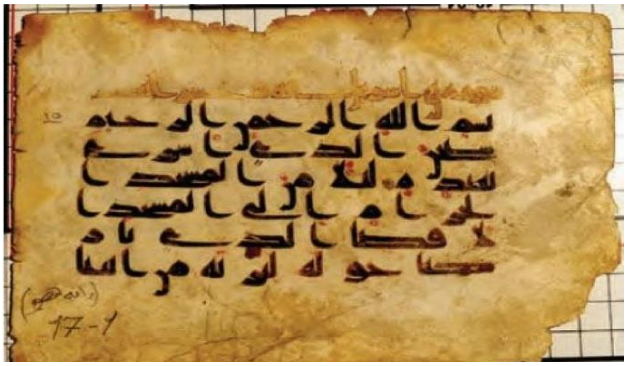


Fig. 3. An Example of Quranic Text.

Historically, with the expansion of the Arab and Muslim empire and the use of Arabic as a global language, it was difficult for many speakers of other languages to distinguish consonants. Thus, the dotting system was introduced in the 12th century [37, 38]. From that time on, Arabic has typically used dots for differentiation. Today, both standard Arabic and colloquial dialects are written using the standard dotting system, as shown in Fig. 4.

However, in the Facebook posts, contrary to usual practice in the standard writing system of Arabic, many users resorted to writing without dotting to circumvent Facebook's algorithms, which are trained to identify, track, and delete content that are classified as offensive and incite hatred in violation of its rules, as shown in Fig. 5.

ا	ا	ـا	ا	alif	ʾ		ʾ
ب	بـ	ـبـ	ب	bāʾ	b		baby
ت	تـ	ـتـ	ت	tāʾ	t		tie†
ث	ثـ	ـثـ	ث	thāʾ	th	th	thin
ج	جـ	ـجـ	ج	jīm	j	dj	job
ح	حـ	ـحـ	ح	ḥāʾ	ḥ	ḥ	‡
خ	خـ	ـخـ	خ	khāʾ	kh	kh	Ger. Buch§
د	د	ـد	د	dāl	d		did†
ذ	ذ	ـذ	ذ	dhāl	dh	dh	then
ر	ر	ـر	ر	rāʾ	r		error (trilled)
ز	ز	ـز	ز	zāʾ	z		zone
س	سـ	ـسـ	س	sīn	s		sand
ش	شـ	ـشـ	ش	shīn	sh	sh, š	shy
ص	صـ	ـصـ	ص	ṣād	ṣ	ṣ	
ض	ضـ	ـضـ	ض	ḍād	ḍ	ḍ	‖
ط	طـ	ـطـ	ط	ṭāʾ	ṭ	ṭ	
ظ	ظـ	ـظـ	ظ	ẓāʾ	ẓ	ẓ	
ع	عـ	ـعـ	ع	ʿayn	ʿ		¶
غ	غـ	ـغـ	غ	ghayn	gh	gh	Fr. rien
ف	فـ	ـفـ	ف	fāʾ	f		fifty
ق	قـ	ـقـ	ق	qāf	q	ķ	ę
ك	كـ	ـكـ	ك	kāf	k	k	kin
ل	لـ	ـلـ	ل	lām	l		lily†
م	مـ	ـمـ	م	mīm	m		maim
ن	نـ	ـنـ	ن	nūn	n		not†
ه	هـ	ـهـ	ه	hāʾ	h		hat
و	و	ـو	و	wāw	w		watch ð
ي	يـ	ـيـ	ي	yāʾ	y		yet □

Fig. 4. The Arabic Alphabet (Source Britannica.com).

Fig. 5. An Example of the use of Arabic without Dots on Facebook.

Among users, to help with this form of writing, different algorithms have been developed to help convert written forms and differentiate them (without using dots) so that their posts are not deleted by Facebook. This has also been used as a way of enabling users to keep their accounts active, rather than being blocked or deleted by Facebook. It was clear that the artificial intelligence algorithms developed by Facebook were not effective in dealing with these non-standard linguistic features of Arabic, which can still be understood by many users even without the dotting system.

V. Analysis and Discussion

Based on the findings of the study, it seems that the artificial intelligence algorithms developed by Facebook for the automatic detection and tracking of hate speech tend not be effective for content in Arabic. This can be attributed mainly to the design of standard automatic detection systems not being appropriate for Arabic content. Arabic, as a Semitic language, has a unique linguistic system that is completely different from Indo-European languages [39]. Today, Arabic is the fifth most widely spoken language globally. It is also ranked fourth in languages used on the Internet [40]. Thus, the linguistic features of Arabic should be considered in the development of artificial intelligence algorithms and automatic detection systems.

The findings of the study agree with the bulk of the related literature in that so far there is no consensus regarding the definition of hate speech. MacAvaney et al. [41] assert that there are disagreements concerning how hate speech should be defined. In our case, it was clear that much of the hate speech in the content identified by Facebook is related to the influence of the religion of Islam. Indeed, many, if not most, hate terms and phrases are taken from or influenced by religious Islamic discourse. For instance, the results showed that posts including the phrases لعنة الله عليهم(May Allah's curse be upon them) and القردة والخنازير (pigs and apes) were tracked and deleted. These phrases were classified by Facebook as inciting hatred against specific groups, namely Christians and Jews. Thus, millions of Facebook users sought to undermine the platform's recognition of these phrases as hate speech by finding ways of deceiving the artificial intelligence algorithms.

In certain interpretations of the Qurʾān, which is believed by Muslims to be the word of God revealed to His prophet Muhammad, the phrase لعنة الله عليهم(May Allah's curse be upon them) is a form of prayer or invocation used to ask Allah

to harm and curse others. According to Ibn manẓūr, those who are thus cursed are rejected by Allah, shunned from his mercy, and hence damned. The verb laʿana means to curse, namely to call upon divine or supernatural power to inflict injury upon somebody. The word laʿana and its derivatives are mentioned 41 times in the Qurʾān, where it is invoked for specific rejected groups of people. For instance, the curse of Allah is invoked upon all those who reject faith in Allah, hypocrites, polytheists, and pagans.

Likewise, the two terms “apes” and “pigs” are used figuratively in the sense of “Carry on behaving like apes and pigs if you want to”, rather than literally [42]. This term of address is given to polytheists. Apes alone are mentioned in the Qur’ān in Chapter/Surat Al-Araf (The Heights) to refer to a specific group of Jews who are blamed by God for their disobedience and breaking the Sabbath by fishing. When the Qur’ān casts blame on Jews, Christians, or the followers of any other religion, it does so specifically on certain people for aberrant behavior, not on the adherents of the religion as a whole [43].

However, contrary to moderate interpretations of the Qur’ān, many phrases have been taken out of context and used to incite hatred against specific groups. Thus, there is a need for religious authorities to point out that such terms and phrases related to particular contexts and specific groups of people, based solely on their lack of belief, transgressions, disobedience, hypocrisy, or aggression, and that it is unacceptable to exploit religious texts, taking such terms and phrases out of context and using them as hate speech on social media.

VI. Conclusion

In recent years, hate speech on social media networks has become a serious challenge for both individuals and institutions. This study aimed to evaluate the performance of artificial intelligence algorithms developed by social media networks for the automatic detection of hate speech. The study was based on evaluating the automatic detection of hate speech in Arabic on Facebook during the 2020 cartoon controversy in France. It can be concluded that automatic detection in Arabic poses a major challenge both for research and social media platforms. This can be attributed to the peculiar linguistic features of Arabic, which are different from those of Western languages. Finally, hate speech in Arabic is greatly influenced by the Muslim religious discourse. Social media posts reproduce verses of Qur’anic text taken out of context and misinterpreting them. Religious organizations and leaders should emphasize that such words and expressions should not be used to disseminate hate or justify hatred and violence.

References

[1] Gagliardone, D. Gal, T. Alves, and G. Martinez, Countering online hate speech. United Nations Educational, Scientific and Cultural Organization, 2015.

[2] C. R. Carlson, Hate Speech. Cambridge, MA: MIT Press, 2021.

[3] F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), 2017, pp. 86-95.

[4] S. Assimakopoulos et al., Online Hate Speech in the European Union: A Discourse-Analytic Perspective. Springer International Publishing, 2017.

[5] J. Golbeck, Online Harassment. Springer International Publishing, 2018.

[6] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in Proceedings of the 28th ACM conference on hypertext and social media, 2017, pp. 85-94.

[7] A. C. Nakaya, Social Media Hate Speech. ReferencePoint Press, 2020.

[8] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in Twitter," Sensors, vol. 19, no. 21, p. 4654, 2019.

[9] S. Modha, T. Mandl, P. Majumder, and D. Patel, "Tracking Hate in Social Media: Evaluation, Challenges and Approaches," SN Computer Science, vol. 1, no. 2, p. 105, 2020/03/28 2020.

[10] E. Walsh. (2021, Oct 17, 2021) Facebook claims it uses AI to identify and remove posts containing hate speech and violence, but the technology doesn't really work. Insider. Available: https://www.businessinsider.com/facebook-ai-doesnt-work-to-remove-hate-speech-and-violence-2021-10.

[11] J. Naughton. (2021, 12 Jun 2021) Big Brother is still watching you and he goes by the name Facebook. The Guardian. Available: https://www.theguardian.com/commentisfree/2021/jun/12/big-brother-is-still-watching-you-and-he-goes-by-the-name-facebook-john-naughton.

[12] N. I. Ariffin and F. Hussain, "The 2020 France Attacks: A Framing Analysis of UK and US Newspapers," International Journal of Modern Trends in Social Sciences, vol. 4, no. 15, pp. 133-146, 2021.

[13] K. Willsher, "Anger spreads in Islamic world after Macron's backing for Muhammad cartoons," in The Guardian, ed, 2020.

[14] A. Omar and B. Deraan, "Towards a Linguistic Stylometric Model for the Authorship Detection in Cybercrime Investigations," International Journal of English Linguistics, vol. 9, no. 5, pp. 182-192, 2019.

[15] R. Moon, Putting Faith in Hate: When Religion Is the Source or Target of Hate Speech. Cambridge: Cambridge University Press, 2018.

[16] A. Guiora and E. A. Park, "Hate speech on social media," Philosophia, vol. 45, no. 3, pp. 957-971, 2017.

[17] A. Omar and B. Deraan, "Cybercrime and authorship detection in very short texts," Opción, vol. 34, pp. 1765-1785, 2019.

[18] A. Nalamothu, O. E. Theses, and D. Center, Abusive and Hate Speech Tweets Detection with Text Generation. Wright State University, 2019.

[19] G. Rehm and T. Declerck, Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings. Springer International Publishing, 2018.

[20] G. Ignatow and R. Mihalcea, An Introduction to Text Mining: Research Design, Data Collection, and Analysis. SAGE Publications, 2017.

[21] A. Choudhary, A. P. Agrawal, R. Logeswaran, and B. Unhelkar, Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2020. Springer Singapore, 2021.

[22] A. T. E. Capozzi et al., "Computational Linguistics Against Hate: Hate Speech Detection and Visualization on Social Media in the “Contro L’Odio” Project," in CLiC-it 2019 Italian Conference on Computational Linguistics, Bari, Italy, 2019: Proceedings of the Sixth Italian Conference on Computational Linguistics.

[23] J. Cobbe, "Algorithmic censorship by social platforms: power and resistance," Philosophy & Technology, pp. 1-28, 2020.

[24] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Computing Surveys (CSUR), vol. 51, no. 4, pp. 1-30, 2018.

[25] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 2017 pp. 1–10: Association for Computational Linguistics.

[26] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," Information fusion, vol. 36, pp. 10-25, 2017.

[27] E. M. Bender and D. T. Langendoen, "Computational Linguistics in Support of Linguistic Theory," Linguistic Issues in Language Technology, vol. 3, no. 2, pp. 1-31, 2010.

[28] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci, "An Italian Twitter Corpus of Hate Speech against Immigrants," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018: European Language Resources Association (ELRA).

[29] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, and C. Bosco, "Hate speech annotation: Analysis of an Italian Twitter corpus," in Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, 2017, vol. 2006.

[30] F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, 2017.

[31] M. M. Bahaa-eddin, HateSpeak in Contemporary Arabic Discourse. . Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 2012.

[32] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: a survey on multilingual corpus," Computer Science and Information Technology, vol. 9, no. 2, pp. 83–100, 2019.

[33] 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.

[34] N. E. Gayar and C. Y. Suen, Computational Linguistics, Speech And Image Processing For Arabic Language. Singapore: World Scientific Publishing Company, 2018.

[35] M. Maroun, "Diacritics and the Resolution of Ambiguity in Reading Arabic," PhD, Department of Psychology, University of Essex, Brighton, 2017.

[36] M. M. Al-Azami, The History of the Quranic Text. Turath Publishing, 2020.

[37] K. Versteegh, Arabic Language. Edinburgh University Press, 2014.

[38] J. F. Healey and G. R. Smith, A Brief Introduction to The Arabic Alphabet. Saqi, 2012.

[39] A. Omar, B. I. Elghayesh, and M. A. M. Kassem, "Authorship Attribution Revisited: The Problem of Flash Fiction A morphological-based Linguistic Stylometry Approach," Arab World English Journal (AWEJ), vol. 10, no. 3, pp. 318-329, 2019.

[40] G. Badaro et al., "A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 18, no. 3, pp. 1-52, 2019.

[41] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," (in eng), PloS one, vol. 14, no. 8, pp. e0221152-e0221152, 2019.

[42] E. M. Badawi and M. Abdel Haleem, Arabic-English Dictionary of Qur'anic Usage. Leiden: Brill, 2008.

[43] M. Abdel Haleem, Exploring the Qur'an: Context and Impact. London/New York: I.B. Tauris, 2017.

AUTHORS' PROFILE

Abdulfattah Omar is an Associate Professor of English Language and Linguistics in the Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University (KSA). Also, he is a standing lecturer of English Language and Linguistics in the Department of English, Faculty of Arts, Port Said University, Egypt. Dr Omar received his PhD degree in computational linguistics in 2010 from Newcastle University, UK. His research interests include computational linguistics, literary computing, digital humanities, discourse analysis, and translation studies.

ORCID: 0000-0002-3618-1750

Mohamed Elarabawy Hashem is an Assistant Professor of English language and Translation Studies at the College of Science and Arts in Tabarjal, Jouf University, KSA & a standing lecturer of English Islamic Studies at the Faculty of Languages and Translation, Al-Azhar University. Cairo. His research interests include Islamic Translation Studies, Semantics, Islamic Discourse Analysis and Hermeneutics.

ORCID: 0000-0001-6818-799X

A Region-based Compression Technique for Medical Image Compression using Principal Component Analysis (PCA)

Sin Ting Lim[1]
Faculty of Engineering and Technology
Multimedia University (MMU)
Melaka, Malaysia

Nurulfajar Bin Abd Manap[2]
Faculty of Electronics and Computer Engineering
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—Region-based compression technique is particularly useful for radiological archiving system as it allows diagnostically important regions to be compressed with near lossless quality while the non-diagnostically important regions (NROI) to be compressed at lossy quality. In this paper, we present a region-based compression technique tailored for MRI brain scans. In the proposed technique termed as automated arbitrary PCA (AAPCA), an automatic segmentation based on brain symmetrical property is used to separate the ROI from the background. The arbitrary-shape ROI is then compressed by block-to-row PCA algorithm (BTRPCA) based on a factorization approach. The ROI is optimally compressed with lower compression rate while the NROI is compressed with higher compression rate. The proposed technique achieves satisfactory segmentation performance. The subjective and objective evaluation performed confirmed that the proposed technique achieves better performance metrics (PSNR and CoC) and higher overall compression rate. The experimental results also demonstrated that the proposed technique is more superior to various state-of-the-art compression methods.**

Keywords*—Principal component analysis; region-of-interest (ROI); automated segmentation; MRI brain scans; region-based compression

I. INTRODUCTION

Since the technological advancement in medical imaging modalities, medical image processing and image analysis have become the important diagnostic aids for medical diagnostics and healthcare. In order for any diagnostic aids to be reliable, the images acquired from imaging modalities need to be of adequate quality and thus requires high amount of resolution. According to the Diagnostic Imaging Dataset Annual Statistics by England (2020) [1], there were 3.8 million MRI test taken in England in between April 2019 to March 2020. The relatively low figure in year 2020 is impacted by the effect of the COVID-19 pandemic but these test images has already summed up to memory storage of as large as Tera bytes per year. The medical images may be required to be saved in PACS and HIS for over thirty years and an efficient compression algorithm is in need to store and archive the images.

Image compression is a process of efficiently coding digital images to reduce the number of bits required in representing an image [2]. Image compression is generally divided into two categories: lossless and lossy. Images compressed by lossless algorithm are perfectly reconstructed but the compression ratio achieved is low. Some common lossless compression methods include Lempel-Ziv-Welch (LZW), Run-Length Encoded (RLE), JPEG Lossless Compression Standard (JPEG-LS), Arithmetic coding and Huffman coding. These methods can only achieve up to 3:1 compression ratio and hence it is not a feasible solution for bulk medical image storage and high speed transmission. Images compressed by lossy algorithm are irreversible but the compression ratio can be ten times higher than the image compressed by lossless algorithm while maintaining good visual quality [3]. Transform coding, vector quantization and predictive coding are three standard methods for lossy image compression [4]. The transform coding techniques, to name a few are – Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA). The recent research in the field of medical image compression involves the wavelet transform are methods such as embedded zerotree wavelet (EZW), set partitioning in hierarchical trees (SPIHT) and embedded block coding with optimized truncation (EBCOT).

Due to the increasing demand for higher compression ratio while keeping the images "visually lossless", research on region-based image compression in medical community is continuously growing. The rationale of utilizing a region-based image compression algorithm is that it exploits prior knowledge on the input image to focus resources on those regions that are significant for making diagnoses. Indeed, medical images are highly structured; for a given imaging modality and given subject matter (i.e. cranial, retina, lung), there are many predictable features in the images that can be taken into account to improve the compression performance [5]. A compressed image should preserve the clinically important features that may be of concern to the radiologist and in most situations, the Non-ROI (NROI) region can be irreversibly compressed with a high compression ratio as the information retained by compression is important only in a contextual sense, helping the viewer to observe the position of the ROI in the original image [6]. In this way, no loss of diagnostically important information can be achieved [7].

There are reports [8-10] that review region-based compression approaches on medical images. Performance evaluation done by Rajkumar and Latte [11] showed that the

PSNR values obtained with region-based compression are not so high compared with those compressed with entire image compression. This is probably due to the use of fractal methods which could lead to insufficient data obtained from the detail image of the wavelet transform. Besides, the use of Huffman coding does not always guarantee a high PSNR value. Region-based image compression was shown to be more suitable for medical images as region-based compression can compress up to 65% while retaining 80% of the original size [12]. In fact, the performance for the region-based image compression could be largely varied based on the ROI selection methods, segmentation goals and compression methods. Previous work in region-based compression mostly focused on compressing different regions with use of different compression schemes while a smaller number of studies on region-based image compression have focused on providing different levels of image quality in different spatial regions. In the work done by Sreenivasulu and Varadarajan [2], wavelet transform and Huffman coding were used to compress the ROI and NROI regions respectively in MRI brain images. Using different segmentation techniques, some researchers [12-13] proposed a Binary Plane Technique to compress the ROI in lossless mode and NROI in lossy mode in MRI brain images. This method offers an advantage in which it is capable of compressing the image both in lossy and lossless mode.

Anastassopoulos and Skodras [14] compared the performance of the general scaling based method and MAXSHIFT method using nephrostogram and reported that MAXSHIFT method achieves better image quality than the general scaling based method. However, both methods may not be suitable for medical images because they do not support lossy-to-lossless compression ROI unless the ROI consists of the whole image [8]. Besides, both methods do not support arbitrary shape for ROI as it is restricted to only support rectangular and circular regions.

The first attempt to compress ROI in medical image using PCA was proposed by Taur and Tau in 1996 [15]. In their research, a simple mean thresholding for blocks of pixels were used to segment the breast tissues. The resulting ROI were either oversegmented or undersegmented and the use of block-by-block PCA algorithm in their work, as proven by our previous research [16], produces very poor image quality. PCA was also employed in study performed in a region-based colour images compression [17] but it was used only to determine the spatio-chromatic information of a colour image so that the existing spatial correlations between the transform coefficients are removed. In region-based compression research performed by Radha [18], foreground of the medical images was identified as the ROI and different compression algorithm such as PCA, EZW, SPIHT and ZTE coding were used to compress the foreground. However, a crude assumption had been made in the research in which ROI is defined as the foreground of the image but a ROI are in fact the area of interest within the foreground. Results comparison performed by Radha show that PCA-based models produce higher compression gain with better PSNR and faster processing speed. The results have also provided a ground base behind the selection of PCA algorithm in this research.

Some of the early studies that applied manual segmentation on region-based compression studies for medical images are reported in [6,19-21]. In general, the reports aforementioned relied on user-defined ROI extracted on a display monitor for different types of medical images. Using the same manual segmentation approach, Seddiki and Guerchi [22] proposed a model that compresses ROI in brain MRI with lossless SPIHT algorithm. Lossy SPIHT compression for NROI has been implemented by Joshi and Rawat [23] and the ROI is selected manually using circular window. Elhannachi et al. [24] extracted the ROI by a rectangular mask and the region is compressed by a lossless EZW coder.

Generally, the existing region-based compression algorithms exhibit a few limitations. Firstly, the ROI is regarded as the whole anatomy without considering the diagnostic values of other portion. Secondly, the ROI is assumed to have a regular shape. Thirdly, most of the automated ROI segmentation is not tailored specifically for brain MRI. Lastly, there is a lack of subjective evaluation towards the efficiency of the schemes. In this context, a region-based compression that addressed the aforementioned limitations is proposed. This work deals with a block-to-row Principal Component Analysis (BTRPCA), which has been reported in our previous work [25] to be effective in the presence of high dimension data, as in image compression.

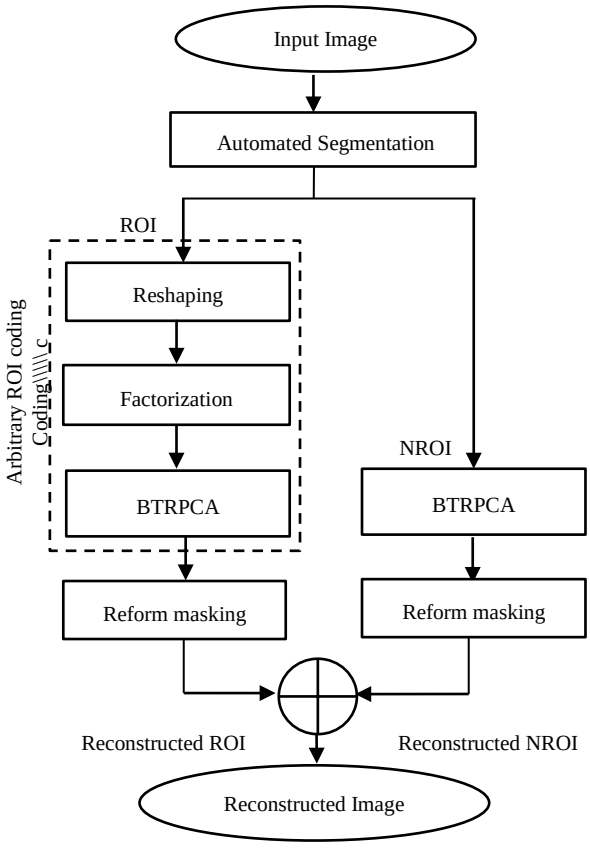


Fig. 1. General Framework for the Proposed AAPCA Algorithm.

The novelty of this work lies in investigating the performance of BTRPCA algorithm coupling with a robust brain segmentation technique in image compression. Fig. 1 depicts the pipeline of our proposed region-based image compression algorithm termed as automated arbitrary PCA (AAPCA). The results of the AAPCA will be subjectively evaluated by a panel of two medical experts and objectively compared with the entire image PCA and state-of-the-art region-based compression algorithms.

The structure of the paper is organized as follows. Section 2 elaborates on AAPCA algorithm for MRI brain images and in Section 3, the objective and subjective evaluation of the proposed methods is verified through experiments. Section 4 concludes the work and finally Section 5 describes future research directions.

II. Materials and Methods

The coding part for the proposed algorithm has been done using Matlab (R2009b), particularly using statistical, wavelet and image processing toolboxes. The statistical analysis has been performed using SPSS Statistics 21.0. The general scaling based method and MAXSHIFT were implemented using JJ2000 version 5.1. To test the robustness of AAPCA towards brain images obtained on different machines using different imaging parameters, selected axial brain images from public datasets namely Radiopaedic [26], Cyprus [27] and Figshare [28] are used as the test images in this study. The test images are in the size of 512 × 512 pixels. The test images are selected so that the brain scans consist of only single ROI, regardless of their acquisition parameters. In each database, the selected images are labelled in a sequence of 1 to 20, depending on the database and the corresponding bit rate (bpp = 1 to 0.0625).

A. Automated Brain Segmentation Technique

The detailed description of our proposed automated brain segmentation technique has been extensively discussed in [29]. The algorithm starts with a robust ellipse fitting technique that extracts the mid-sagittal plane (MSP) of the brain. This shape-based method enjoys robustness towards low signal to noise brain images by assuming the skull of the head to be in elliptical shape. Once the MSP has been successfully extracted, the brain images will be tilted either to the left or right to ensure that the brain images can be equally dissected into left and right hemisphere. The Absolute Difference Algorithm (ADM) that involves a series of absolute summation and absolute difference operation are then performed on the left and right hemispheres. In this work, the ADM algorithm was further improved to increase the segmentation rate for smaller ROI. Fig. 2 shows the step-by-step image manipulation for right hemisphere and its flipped-left hemisphere. Although it is not shown, the same operations are performed for left hemisphere and its flipped-right hemisphere to obtain the largest connected component (LCC) in the hemisphere. This method delineates and highlights the differences in both hemispheres.

With this approach, no a prior knowledge is needed on whether the ROI is located at the left hemisphere or right hemisphere and it is relatively simpler to compute than the ADM method proposed by Liu et al. [30]. Thresholding operation is then applied to select only the high intensity region in the image. The threshold value has been determined based on Otsu's thresholding where this method finds a threshold value between the peaks of a histogram. Once each hemisphere is left with the LCC, they will be merged again to form a whole brain image. This is when the morphology fill operator will be applied so that the less significant regions are removed by filling in holes and small pits from the edge. The resulting ROI will then be superimposed on the original brain image and this completes the final segmentation of the ROI. The output image is then ready for subsequent region-based compression.

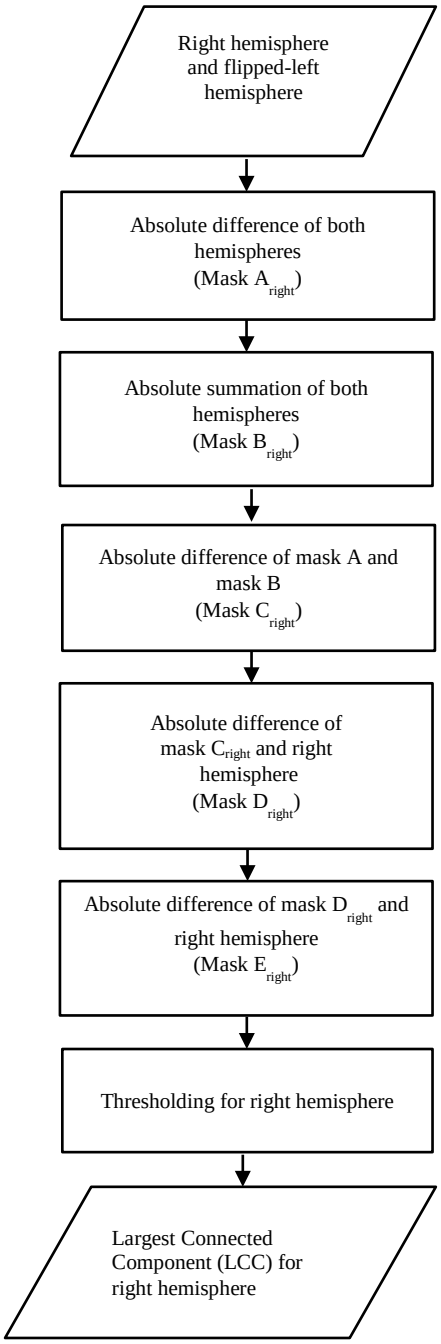


Fig. 2. Proposed ADM for Right Hemisphere.

B. Arbitrary ROI Coding (ARC)

The proposed ARC is a compression technique where arbitrary-shape ROI are explicitly defined, reshaped, factorized and compressed in a near lossless way whereas the NROI are compressed in a lossy manner. In the present work, the proposed ARC is mainly divided into two parts: (1) Reshaping and factorization of ROI and (2) Selective compression of the ROI and NROI with BTRPCA.

Since the segmented ROI is in arbitrary shape, the regions will be reshaped in this step to form a matrix. In this step, the pixel values of the arbitrary-shape ROI will first be converted into a row, resulting in a vector of $1\times q$ as shown in equation (1) below:

$$R_r = \begin{bmatrix} R(0,0) & R(0,1) & \dots & R(0,q) \end{bmatrix}_{(1\times q)} \tag{1}$$

where q represents the total number of pixel values contained within the regions.

With the objective of constructing a matrix from a row vector, the divisors of q can be obtained using factorization. Assume that the total number of divisors of q is found to be t, the row and column of the new transformed matrix will be selected based on the following statements given $i(0) \le i(t/2) \le i(t)$ [31]:

1) If *t* is even, then *i* is selected from position of $t/2$ from the list of divisors.

2) If *t* is odd, then *i* is selected from position of $(t+1)/2$ from the list of divisors.

where i is the row and q/i the column of the matrix.

$$R_f(x,y) = \begin{bmatrix} R(0,0) & R(0,1) & \dots & R(0,j-1) \\ R(1,0) & R(1,1) & \dots & R(1,j-1) \\ \vdots & \vdots & \vdots & \vdots \\ R(i-1,0) & R(i-1,1) & \vdots & R(i-1,j-1) \end{bmatrix}_{(i\times j)} \tag{2}$$

The formulated matrix that carries the ROI information by now has the size of $(i\times j)$ as shown in the equation (2) that is ready to be compressed using block-to-row algorithm. For instance, the segmented arbitrary ROI contains a total of 88 pixels. By factorization, the divisors of q are found to be 1, 2, 4, 22, 44 and 88. The total number of divisor t is thus 6. Since t is an even number, i is selected from the 3rd position from the list of divisors and the number of row in the new matrix is 4. The selected arbitrary ROI will then be reshaped and compressed based on the size of 4 × 22 where 22 being the number of 88 divide by 4.

The ROI matrix $R_f(x,y)$ and NROI matrix $NR_f(x,y)$ will be partitioned into $n\times n$ blocks and the mean-subtracted transformed matrix shown in equation (3) will consist of the mean row vector of each block:

$$\overline{D_{roi}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \overline{x_3} \\ \dots \\ \overline{x_b} \end{bmatrix}_{(b\times n^2)} \tag{3}$$

If the size of the medical image is not a multiple of n, zero paddings will be performed by adding zeros at the borders. The number of blocks b can be determined from block division with the following equation:

$$b = \frac{N\times M}{n^2} \tag{4}$$

where N is the number of row and M is the number of columns for the original image.

The resulting feature matrix, Vroi that contains only the chosen k principal components are given as:

$$V_{roi} = \begin{bmatrix} \lambda_1, \lambda_2, \lambda_3, \dots, \lambda_j \end{bmatrix}_{(n^2\times k)} \tag{5}$$

The ROI data is now compressed as in equation (6):

$$Y_{roi} = \begin{bmatrix} V_{roi}{}^T * \overline{D}_{roi}{}^T \end{bmatrix}_{(k\times b)} \tag{6}$$

Similarly, the NROI region $NR_f(x,y)$ will also be fed into block-to-row algorithm where the NROI compressed data is obtained with the reduced feature matrix V_{nroi} and mean transformed matrix $\overline{D}_{nroi}$ as shown in equation (7). K is the chosen principal components and B is the number of blocks.

$$Y_{nroi} = \begin{bmatrix} V_{nroi}{}^T * \overline{D}_{nroi}{}^T \end{bmatrix}_{(K\times B)} \tag{7}$$

At this stage, the dimensionality of the ROI data has been reduced from n to p column. Since the region-based algorithm divides an image into two regions, the compression ratio is related to the size of the ROI and NROI. The compression ratio of the ROI is therefore defined as [32]:

$$\begin{aligned} CR_{ROI} &= \left(1 - \frac{Size\ of\ compressed\ ROI\ data}{Size\ of\ original\ ROI\ data}\right) \times \frac{total\ pixels\ in\ ROI}{Size\ of\ original\ image} \\ &= \left(1 - \frac{|k|\times b}{i\times j}\right)\left(\frac{i\times j}{M\times N}\right) \\ &= \left(\frac{ij - |k|b}{MN}\right) \end{aligned} \tag{8}$$

No compression occurs if the number of principal components equals to the square of block size ($k = n^2$). The compression ratio of the NROI is defined as:

$$CR_{NROI} = \left(1 - \frac{Size\ of\ compressed\ NROI\ data}{Size\ of\ original\ NROI\ data}\right) \times \frac{total\ pixels\ in\ NROI}{Size\ of\ original\ image}$$
$$= \left(1 - \frac{|K| \times B}{I \times J}\right)\left(\frac{I \times J}{M \times N}\right)$$
$$= \left(\frac{IJ - |K| B}{MN}\right) \quad (9)$$

The following steps explain the procedure of the compression scheme:

Step 1: The ROI is encoded first on a high priority basis followed by the NROI on a low priority.

Step 2: Restructure the ROI into a matrix in which the size p × q is determined by the divisors obtained.

Step 3: Divide the ROI into a set of blocks, $s_{ij}\left(i = 1, \ldots, \left[\frac{p}{n_1}\right], j = 1, \ldots, \left[\frac{q}{n_1}\right]\right)$, where the value of n1 used in the current experiment is 8.

Step 4: Perform zero padding to the ROI if the size is not in the multiple of n1.

Step 5: Compress the ROI selectively using block-based PCA algorithm with low CR and high bpp as desired i.e. $CR_{\min} < CR_{ROI} < CR_{\max}$ where the range of the CRmin and CRmax vary according to the desired quality of reconstruction.

Step 6: Divide the NROI into a set of blocks, $S_{ij}\left(i = 1, \ldots, \left[\frac{M}{n_2}\right], j = 1, \ldots, \left[\frac{N}{n_2}\right]\right)$, where the value of n2 used in the current experiment is 8 and 16.

Step 7: Perform zero padding to the NROI if the size is not in the multiple of n2.

Step 8: Compress the NROI selectively with the block-based PCA algorithm with high CR and low bpp as desired i.e. $CR_{whole} = CR_{ROI} + CR_{NROI}$ and $CR_{\min} < CR_{NROI} < CR_{\max}$ where the range of the CRmin and CRmax vary according to the desired quality of reconstruction.

Step 9: Compare and perform pixel by pixel analysis for original and reconstructed image using CoC analysis. If the CoC is not within satisfactory range between 0.9 and 1.0, adjust the compression rate on ROI and NROI and go to step 8.

III. Results and Discussion

In this section, we report the quantitative analysis for the segmentation technique, the subjective experiment conducted to assess the quality of images compressed with proposed method and entire image method, and the objective metrics resuls. We also present and compare the proposed AAPCA with the mainstream compression methods using default parameters except as noted. We also present the simulated results for region-based algorithm proposed by Sreenivasulu and Varadarajan [2] (i.e. SV algorithm).

A. Segmentation Performance

The overall ROI segmentation performance are evaluated using segmentation score *S* as shown in equation (10) and the computation time in seconds. The segmentation score is used to evaluate the effectiveness or clustering operation of a segmentation algorithm and it is mathematically represented as [33]:

$$S = \sum_{n=1}^{c} \frac{f(x,y)_n \cap f(x,y)_{refn}}{f(x,y)_n \cup f(x,y)_{refn}} \quad (10)$$

where $f(x,y)_n$ represents the set of pixels belonging to the nth class found by the algorithm while $f(x,y)_n$ represents the set of pixels belonging to the nth class in the ground truth segmented image. One participant was requested to use a HUION HS64 drawing tablet to draw the contour of the lesions on the test images displayed directly on a computer screen and the drawn results will be served as the ground truth to compare with the ROI segmented by the proposed method and segmentation algorithm proposed by Liu et al. [30].

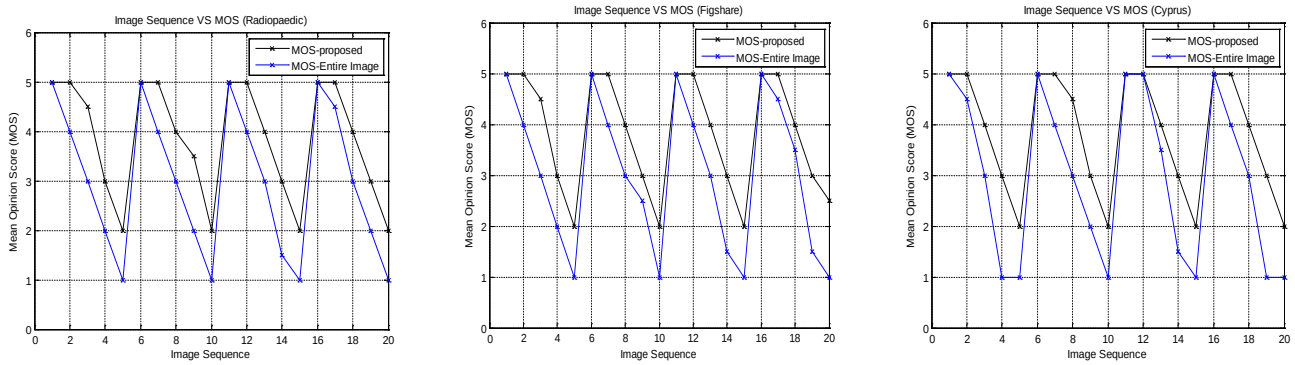
The proposed segmentation algorithm on 67 MRI brain scans achieves an average of 0.7414 ± 0.086 segmentation scores while the algorithm proposed by Liu et al. achieves higher scores of 0.7823 ± 0.068. Although the increase in mean segmentation score is 0.0409 (5.5%), Liu et al. algorithm took an average of 120.9627 seconds to process an image, as compared to the proposed algorithm that took 8.4294 seconds. Since the proposed segmentation algorithm is to work with a cascading compression algorithm, the whole infrastructure is aimed to be computational efficient. Hence the proposed segmentation has significantly achieved shorter computation time in the study.

B. Subjective Evaluation

To assess image quality to diagnostic utility, image quality for output images compressed at various bpp were evaluated subjectively by a panel of one opthalmologist and one radiologist (A and B). Each panel was presented independently with test images arranged randomly and anonymously. Four MRI brain images from three dataset compressed at five different compression ratio (over the range of bpp = 0.0625 to 1.0) using entire image PCA and the proposed arbitrary methods sum up to a total of 120 test images. These images were shown on the computer screen and the panels were asked to rank the images based on the criteria as shown in Table I. The panels were asked to rank the images in two sessions held at least two weeks apart. Each session consisted of 60 compressed images with randomized order.

TABLE I. MOS for Subjective Evaluation

MOS	Description	Comments
5	Excellent (Imperceptible Distortion)	Useful for Diagnosis Purposes
4	Good (Perceptible Distortion but not Annoying)	Useful for Diagnosis Purposes
3	Fair (Slightly Annoying but acceptable)	Useful for Diagnosis Purposes
2	Bad (Annoying)	Not Useful for Diagnosis Purposes
1	Very bad (Very Annoying)	Not Useful for Diagnosis Purposes



(a) (b) (c)

Fig. 3. Image Sequence Versus MOS for All Tested Image from Three Datasets (a) Radiopaedic (b) Figshare (c) Cyprus.

Fig. 3 shows the MOS for all tested images for proposed algorithm and its traditional counterpart. As expected, the resulted MOS decreases for images compressed at lower bit rate on our datasets. However, it is observed that the MOS values obtained in the proposed method are higher than that of entire image method except for bpp = 1.0 where image sequences are 1, 6, 11 and 16. For images compressed at bpp = 1.0, both methods unanimously correspond to "Imperceptible distortion" scoring level. It can also be observed that the MOS values for proposed method and entire image algorithm are closer at higher bpp but the MOS decreases more drastically in the case of entire image method as bpp reduces from 1.0 to 0.0625. As shown in the contigency table for both reviewers in Table II, there is a 88.3% of inter-reviewer agreement of scoring for a total of 120 compressed images. The intra-class correlation coefficient for both reviewers is 0.970 [0.957, 0.979] and the mean difference in scores is 0.05 [0.72 -0.62]. The normality test using Komogorov-Sminov and Shapiro-Wilk showed that the distribution of scores are non-normal hence the differences in scores between two methods were compared using a non-parametric test called Wilcoxon's match-pairs signed rank test. The Wilcoxon signed rank test shows that the MOS scores for proposed method at bpp = 0.0625, 0.125, 0.25 and 0.5 differ significantly from the MOS scores for entire image method at p = 0.001, p = 0.002, p = 0.002 and p = 0.001 respectively for a two-tailed test. Wilcoxon signed rank tests did not yield any significant differences between the MOS scores for proposed method and the MOS scores for entire image method at bpp = 1.0 (p = 1.0).

C. Objective Evaluation

This study includes two image quality metrics PSNR and Correlation coefficient (CoC) as shown in equation (12) and (13). Suppose that X is the original image and Y is the compressed/reconstructed image with size of m × n, where X_{ij} and Y_{ij} the values of the ith and jth pixels in X and Y respectively, the MSE is the cumulative squared error between the original and the compressed image:

$$MSE(\mathbf{X},\mathbf{Y}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-Y_{ij})^2 \tag{11}$$

TABLE II. CONTINGENCY TABLE OF SCORES

	Panel B					
Panel A	Score	5	4	3	2	1
	5	36 (30.0%)	3 (2.5%)	-	-	-
	4	2 (1.67%)	18 (15.0%)	2 (1.67%)	-	-
	3	-	1 (0.83%)	22 (18.3%)	1 (0.83%)	-
	2	-	-	1 (0.83%)	18 (15.0%)	4 (3.33%)
	1	-	-	-	-	12 (10.0%)

In the literature of image compression, MSE is often converted into the PSNR measure:

$$PSNR = 10\log_{10}\frac{L^2}{MSE} = 10\log_{10}\frac{(2^B-1)}{MSE} \tag{12}$$

where L is the dynamic range of allowable image pixel intensities and *B* is number of bits that represent a pixel. PSNR is measured in the unit of decibel (dB) and the metric provides a straightforward notion related to the image fidelity - the higher the PSNR value, the higher the image fidelity and vice versa. The second performance metric is CoC that suggests how closely the reconstructed image is correlated with an original image, on a scale of 0-1. The closer the value of CoC to 1, the higher the correlation of the compressed image to the original image is. The CoC is defined as:

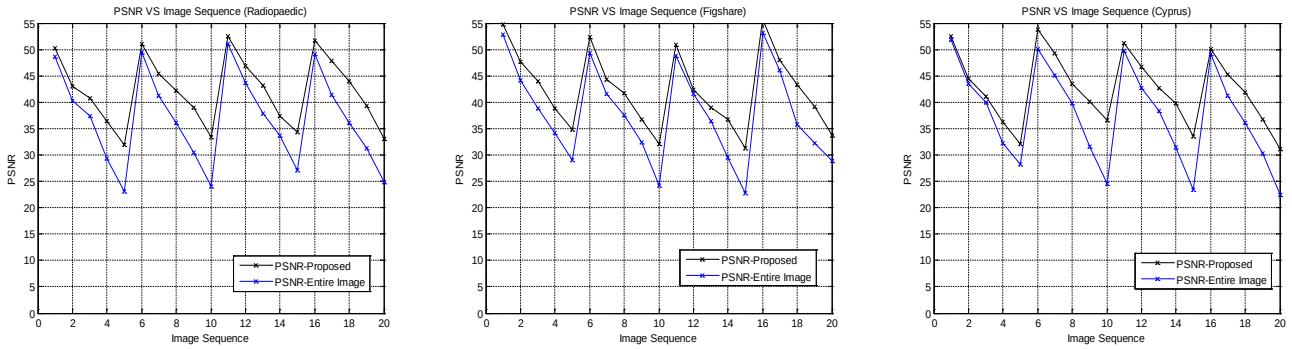
$$CoC(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}Y_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}^{\,2}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}Y_{ij}^{\,2}}} \tag{13}$$

The values of PSNR and CoC are plotted against the image sequence as shown in Fig. 4 and Fig.5 respectively for our datasets. It can be observed that across all graphs, the AAPCA clearly outperforms the entire image algorithm in terms of PSNR and CoC. The resulting trend somewhat echoed the

subjective evaluation performance achieved by the proposed method. The average PSNR of AAPCA is compared with JPEG, JPEG2000, EZW, SPIHT and entire image PCA that are applied to the whole image. As shown in Fig. 7, the proposed method performs better than JPEG, EZW and entire image PCA at all tested bit rate. Statistical results demonstrate that the mean PSNR for AAPCA increased significantly with the mean PSNR for the EZW (30.4930 vs. 47.9696, $p < 0.001$) and JPEG (40.4758 vs. 47.9696, $p < 0.001$). It is interesting to learn that while EZW performance fell behind the other compression methods, the PSNR performance for entire image PCA and JPEG were close to each other. The AAPCA performs slightly inferior to JPEG2000 at low bpp. The rate-distortion performance for AAPCA is equivalent to that of SPIHT at high bpp but inferior to SPIHT at low bpp. This can probably be explained by the fact that progressive transmission in SPIHT reduces the MSE distortion more significantly for every bit-plane sent even though the image is compressed at high compression ratio.

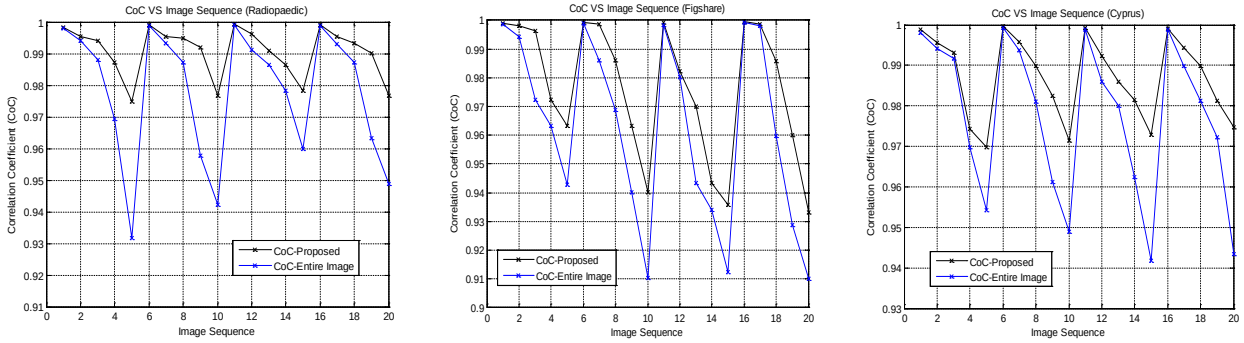
The AAPCA is also compared with the four existing region-based methods and it is observed from Fig. 8 that the proposed method achieves higher PSNR than EBCOT and SV algorithm. The mean PSNR for the AAPCA increased significantly with the mean PSNR for the EBCOT (37.0275 vs. 47.9696, $p < 0.001$) and SV algorithm (30.9163 vs. 47.9696, $p < 0.001$). However the average PSNR for AAPCA is seen to be lower than MAXSHIFT towards the higher end (above 0.8 bpp) and lower end (below 0.125 bpp) of the bpp as shown in Fig. 8. Similarly the PSNR of AAPCA is higher than the general scaling based method except at higher end (above 0.8 bpp) and lower end (below 0.125 bpp) of the bpp. A reason for the drop of PSNR compare to general scaling based method and MAXSHIFT is due to the block-based nature of the block-based PCA algorithm. It should also be noted that while AAPCA needs to encode the shape information of the ROI, it is not necessary with the MAXSHIFT method, enabling lower computational cost.

The original and reconstructed image output of AAPCA for part of the test images at bpp = 1.0, bpp = 0.25 and bpp = 0.0625 are provided in Fig. 6. The ROI automatically segmented using the AAPCA is also included in the figures to illustrate the region selected and extracted by the algorithm. The NROI (not shown on the figures) is the image region void of the ROI region. The same common observation made across all three databases is that reconstructed images suffer little or no visual distortion at bpp = 1.0 and the visual quality is maintained even at lower bit rate, bpp = 0.25. Notice that there are no visible blocking and unnatural noise artifacts. However, the image quality of the reconstructed images compressed at bpp = 0.0625 is slightly deteriorated and the images are seen to be impaired to a certain extent, preserving only the image quality of the ROI. The impairment, if noticeable, are blocking artifacts usually exhibited at the ROI edges and blurring. These artifacts are the reason the proposed method loses PSNR compared to SPIHT, the general scaling based method and MAXSHIFT at low bit rate. Another common observation that can be made is that the arbitrary-shape ROI are faithfully represented using AAPCA compressed at different bit rates.



(a) (b) (c)

Fig. 4. Image Sequence Versus PSNR for All Tested Image from Three Datasets (a) Radiopaedic (b) Figshare (c) Cyprus.



(a) (b) (c)

Fig. 5. Image Sequence Versus CoC for All Tested Image from Three Datasets (a) Radiopaedic (b) Figshare (c) Cyprus.

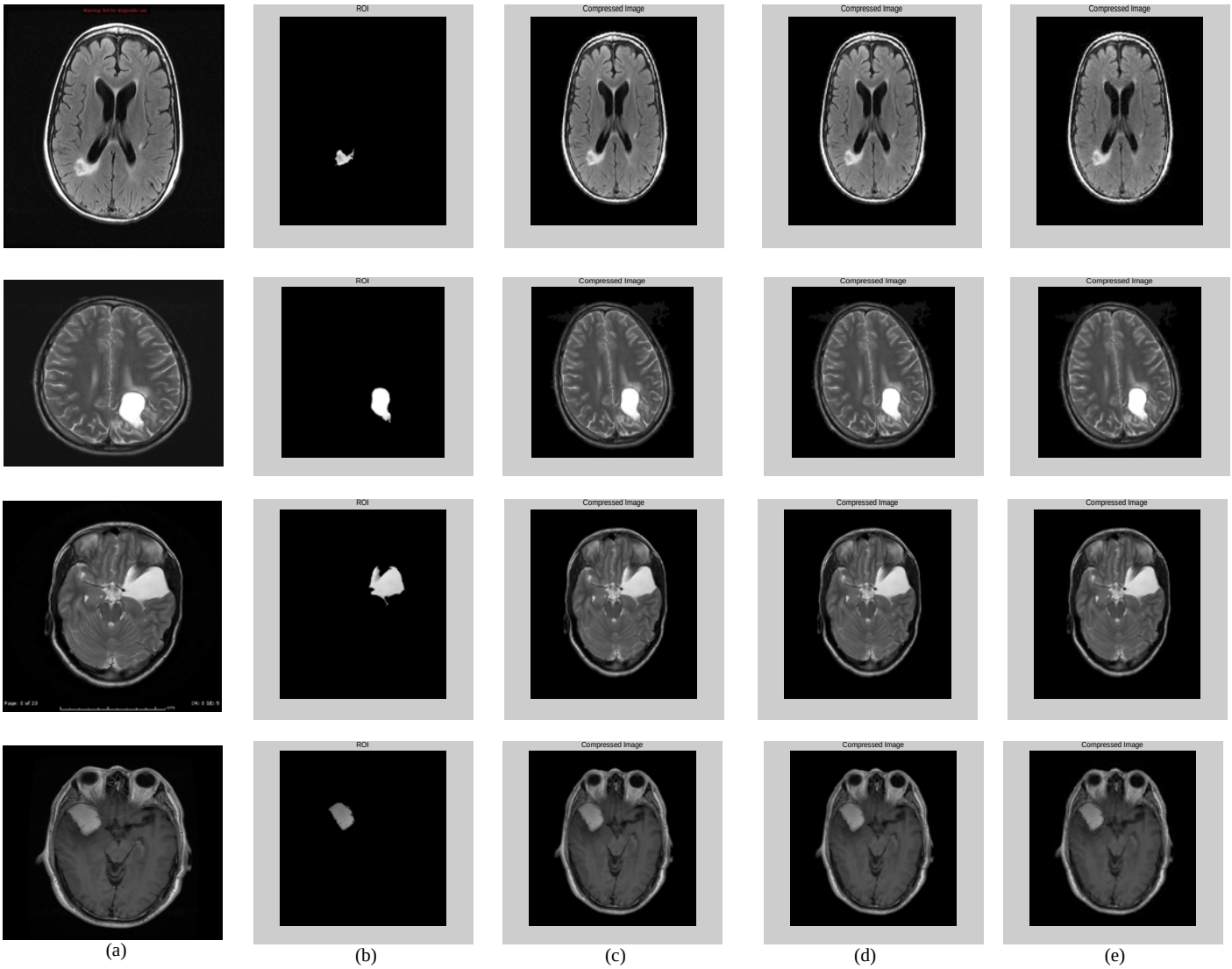


Fig. 6. MRI Brain Images Reconstructed with the Proposed Algorithm at different bpp and CR (a) Original Images (b) Extracted ROI using AAPCA (c) Reconstructed Images at bpp = 1.00 (d) Reconstructed Images at bpp = 0.25 (e) Reconstructed Images at bpp = 0.0625.

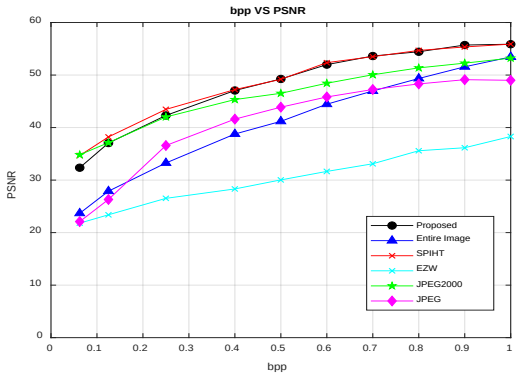


Fig. 7. Graphical Presentation of bpp vs. Average PSNR for JPEG, JPEG2000, EZW, SPIHT, Entire Image PCA and Proposed AAPCA.

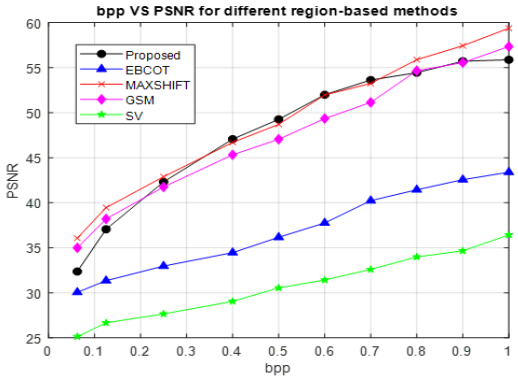


Fig. 8. Comparison of Average PSNR between Proposed AAPCA and the Existing Region-based Compression Algorithms.

IV. Conclusion

In this work, to provide solutions for an efficient region-based image compression, we presented an automated segmentation/compression algorithm termed as automated arbitrary Principal Component Analysis (AAPCA), which takes into account of arbitrary shape ROI using principal component analysis as the core compression method. The experimental results demonstrated that this technique is applicable to axial brain scans for which the quality of an arbitrary shape region is desired to be lossless and the brain segmentation is to be automated. We successfully applied AAPCA in the framework of MRI brain image compression from three public databases where the ROI is the lesion.

The objective and subjective evaluation confirmed that the proposed method is capable of extending beyond the compression limits of conventional PCA algorithm. Statistical analysis shows that AAPCA outperforms JPEG, EZW, EBCOT, and technique proposed by [2] both in CR and reconstruction quality. No image noise and blockiness are observed at bpp as low as 0.25. The only limitation is that there is possibility of loss of information at the edge of the ROI at high compression rate. However one may argue that the compression parameter of the image can be optimally adjusted so that the edge of the ROI can be highly preserved without compromising the bit rate. Although the proposed algorithm is markedly tailored to MRI brain images, similar algorithms can be devised for other image modality or anatomy of interest. The main objective is to focus resources on a given medical image modality and exploit the knowledge of its invariant features.

V. Future Work

While the results gathered in the study are promising, there is room for improvement. One topic that deserves further investigation is that AAPCA is a 2D model. In order to better segment the ROI, the algorithm can be implemented in 3D environment to enable better analysis and detection of ROI in 3D since 3D based models are less susceptible to the disorientation of MSP. On top of that, implementing segmentation in 3D environment allows the segmentation/compression algorithm to compress the whole volume of ROI. Continuation of this study might include suppressing the blocking artifacts with the use of a post-processing algorithm such as a deblocking filter.

Future work can also include the extraction of multiple ROI regions prior to compression for example in the case of metastatic brain tumors. One may also choose the best suited region-based compression technique for all kinds of telemedicine archiving system based on a specific application. It would be interesting to predict visual quality of other compression technique. The current proposed method is believed to be a probable tool for future research focusing on medical images where arbitrary ROI coding is of concern in multimedia application, and even some telecommunication applications. Along with its value for the specific application, the presented results in this thesis reveal the fact that segmentation/compression published thus far can be improved considerably. The benefits reaped are significant by developing more powerful segmentation technique.

Acknowledgment

This study was supported in part by grant number FRGS/2012/FKEKK/TK02/03/1/F0012 and PRGS/2012/ TK02/FKEKK/01/1/T0002 from the Ministry of Higher Education of Malaysia.

References

[1] National Health Service England, 2020. Diagnostic Imaging Dataset Annual Statistical Release 2019/20.

[2] P. Sreenivasulu and S. Varadarajan, “An efficient lossless ROI image compression using wavelet-based modified region growing algorithm,” Int J Intell Syst, vol. 29, no. 3, pp.1063-1078, 2020.

[3] R. Kaur and P. Choudhary, “A review of image compression techniques,” Int J Comput Appl, vol. 142, no. 1, pp.8-11, 2016.

[4] Y. Pourasad and F. Cavallaro, “A novel image processing approach to enhancement and compression of x-ray images,” Int J Environ Res Public Health, vol. 18, no. 13, pp.6724, 2021.

[5] G. Poggi and R. A. Olshen, “Pruned tree-structured vector quantization of medical images with segmentation and improved prediction,” IEEE Trans Image Process, vol. 4, no. 6, pp.734-742, 1995.

[6] A. Vlaicu, S. Lungu, N. Crisan and S. Persa, “New compression techniques for storage and transmission of 2D and 3D medical images,” Proceedings of SPIE, Visual Communications and Image Processing '95, Taipei, Taiwan, 21 April 1995. SPIE Digital Library.

[7] M. Kim, Y. Cho, D. Kim and N. Ha, “Compression of medical images with Regions of Interest (ROIs),” Proceedings of SPIE, Visual Communications and Image Processing '95, Taipei, Taiwan, 21 April 1995. SPIE Digital Library.

[8] C. Doukas and I. Maglogiannis, “Region of interest coding techniques for medical image compression,” IEEE Eng Med Biol, vol. 26, no. 5, pp.29 – 35, 2007.

[9] R. L. Verma, D. Ojha, M. P. Gupta and M. M. Gupta, “A review report on ROI based encoding an effective technique of compression for medical imaging,” Int J Sci Eng, vol. 8, no. 4, pp.2094-2099, 2013.

[10] R. Kaur and R. Rani, “ROI and Non-ROI based medical image compression techniques: a survey and comparative review,” International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 15-17 December 2018. IEEE.

[11] T. M. P. Rajkumar and M. V. Latte, “Adaptive thresholding based medical image compression technique using Haar wavelet based listless SPECK encoder and artificial neural network,” J Med Imaging & Health Infor, vol. 5, no. 2, pp.223-234, 2015.

[12] B. P. S. Kumar and K. V. Ramanaiah,“Region based image compression with deep learning and binary plane difference methods. In Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies. Lecture Notes in Electrical Engineering, (V. Gunjan, , S. Senatore, A. Kumar, X. Z. Gao and S. Merugu, 1st ed.), 643, pp.325-332. Singapore: Springer.2020.

[13] V. Krishna, and V. P. C. Rao, “Region based medical image compression with binary plane coding,” J Eng Appl Sci., vol. 12, no. 17, pp.5124-5128, 2017.

[14] G. Anastassopoulos and A. Skodras, “JPEG 2000 ROI coding in medical imaging applications,” Proceedings of the 2nd IASTED International Conference on Visualisation, Imaging and Image Processing, Marbella, Spain, 9-12 September 2002. ACTA Press.

[15] J. S. Taur and C. W. Tao, “Medical image compression using principal component analysis,” Proceedings of the International Conference on Image Processing, Lausanne, Switzerland, 16-19 September 1996. IEEE.

[16] S. T. Lim, F. W. D. Yap and N. A. Manap, “Medical image compression using block-based PCA algorithm,” International Conference on Computer, Communication and Control Technology, Langkawi, Malaysia, 2-4 September 2014. IEEE.

[17] D. Carevic and T. Caelli, “Region-based coding of color images using Karhunen–Loeve transform,” CVGIP-Graph Model IM, vol. 59, pp.27-38, 1997.

[18] V. Radha, "A comparative study on ROI-based lossy compression techniques for compressing medical images," Proceedings of the International Conference on World Congress on Engineering and Computer Science, San Francisco, USA, 19-21 October 2011. International Association of Engineers (IAENG).

[19] C. C. Sim, W. C. Wong and K. Ong, "Segmented approach for lossless compression of medical images," Proceedings of IEEE Singapore International Conference on Networks/International Conference on Information Engineering, Singapore, 6-11 September 1993. IEEE.

[20] J. Ström and P. C. Cosman, "Medical image compression with lossless regions of interest," Signal Process, vol. 59, pp.155-171, 1997.

[21] M. Kim, Y. Cho, D. Kim and N. Ha, "Compression of medical images with Regions of Interest (ROIs)," Proceedings of SPIE, Visual Communications and Image Processing '95, Taipei, Taiwan, 21 April 1995. SPIE Digital Library.

[22] A. Seddiki and D. Guerchi, "Medical image compression by region of interest based on SPIHT and global thresholding using Huffman coding," Recent Advances in Electrical Engineering and Educational Technologies, pp.235-238, 2014.

[23] P. V. Joshi and C. D. Rawat, "Hybrid compression for medical images using SPIHT," Int J Curr Eng Sci Res, vol. 3, no. 7, pp.62-69, 2016.

[24] S. A. Elhannachi, N. Benamarane and T. Abdelmalik, "Adaptive medical image compression based on lossy and lossless embedded zerotree methods," Int J Ind Syst Eng, vol. 13, pp.40-56, 2017.

[25] S. T. Lim, F. W. D. Yap and N. A. Manap, "A GUI system for region-based image compression using principal component analysis," International Conference on Computational Science and Technology *(ICCST 2014)*, Kota Kinabalu, Malaysia, 27-28 August 2014. IEEE.

[26] Radiopaedic, 2005. Cases.

[27] Cheng, J., 2017. *Brain Tumor Dataset. Figshare*.

[28] E-Health Lab, Department of Computer Science, University of Cyprus, 2011. MRI Lesion Segmentation in Multiple Sclerosis Database.

[29] S. T. Lim, F. W. D. Yap and N. A. Manap, "Automated ROI-based compression on brain images using principal component analysis, " J Eng Appl Sci, vol. 13, no. 14, pp.5967-5970, 2018.

[30] S. X. Liu, C. Imielinska, A. Laine, W. S. Millar, E. S. Connolly and A. L. D'Ambrosio, "Asymmetry analysis in rodent cerebral ischemia models." Acad Radiol, vol. 15, no. 9, pp.1181-1197, 2008.

[31] S. T. Lim, F. W. D. Yap and N. A. Manap, "A novel approach for arbitrary-shape ROI compression of medical images using Principal Component Analysis (PCA)," Trends Appl Sci Res, vol. 10, no. 1, pp. 68–76, 2015.

[32] N. B. Bahadure, A. K. Ray and H. P. Thethi, "Comparative approach of MRI-based brain tumor segmentation and classification using genetic algorithm," J Digit Imaging, 31, pp.477-489, 2018.

[33] M. Gong, Y. Liang, J. Shi, W. Ma and J. Ma, "Fuzzy C-Means clustering with local information and kernel metric for image segmentation," IEEE Trans Image Process, vol. 2, no. 22, pp.573-584, 2013.

Identify Discriminatory Factors of Traffic Accidental Fatal Subtypes using Machine Learning Techniques

W.Z. Loskor[1]
Department of Science and Humanities
Bangladesh Army International University of Science and Technology, Cumilla-3501, Bangladesh

Sharif Ahamed[2]
Department of Computer Science and Engineering
Gono Bishwabidyalay, Dhaka-1344
Bangladesh

***Abstract*—In today's world, traffic accidents are one of the main reasons of mortality and long-term injury. Bangladesh is no exception in this case. Several vehicle accidents each year have become an everyday occurrence in Bangladesh. Bangladesh's largest highway, the Dhaka-Banglabandha National Highway, has a significant number of accidents each year. In this work, we gathered accident data from the Dhaka-Banglabandha highway over an eight-year period and attempted to determine the subtypes present in this dataset. Then we tested with various classification algorithms to see which ones performed the best at classifying accident subtypes. To describe the discriminatory factors among the subtypes, we also used an interpretable model. This experiment gives essential information on traffic accidents and so helps in the development of policies to reduce road traffic collisions on Bangladesh's Dhaka-Banglabandha National Highway.**

***Keywords*—Traffic accident; clustering analysis; machine learning; feature selection; classification; discriminatory factors**

I. Introduction

Traffic accidents have become one of the leading causes of loss of life and property. The likelihood of traffic accidents is increasing as the number of vehicles and roads increases. In 2020, total of 4,891 vehicle accidents in Bangladesh killed 6,686 people and wounded 8,600 others [1]. As a result, 18 individuals died in traffic accidents each day across the country. In its yearly road accident observing report for 2020, Bangladesh Passengers Welfare Association (BPWA) disclosed these data. According to the Accident Research Institute of Bangladesh University of Engineering and Technology (BUET), 56,987 individuals have perished in 58,208 vehicle accidents in Bangladesh in the last two decades. Many researchers have examined road accident datasets and used various machine learning methods to predict the risk of an accident; some of their findings are summarized in the Literature Review section. All of the research efforts on these datasets are aimed at classifying the risk of a car accident. Many organizations exist in many countries around the world to maintain road safety in order to reduce the threat of fatal road traffic accidents. Researchers, more-over, used a variety of techniques, particularly statistics methods, to define the reasons of traffic road accidents through a historical path traffic road dataset. Using various data mining tools and techniques, the data mineworkers investigated different parameters or variables for the reasons of traffic accidents besides diver behaviors. A lot of researchers used to expend a significant amount of time attempting to find the greatest performing data mining procedure for mining the traffic road accidents dataset.

In this research, we collected traffic accident data in the DBH from the Accident Research Institute (ARI), BUET, from 2007 to 2015, and we only used fatal data records. This research aims to identify only accident fatal subtypes and identify important discriminant features that will assist authorities in better understanding accident risks.

The rest of this paper is as follows. Section II is dedicated to related activity. Section III discusses traffic accident data analysis and methodology. The findings of the experiments are contained in Section IV. Finally, Section V summarizes the work's findings.

II. Literature Review

Several studies have been launched to investigate traffic accident data using various approaches. In 2021, M. Bobermin et al. suggested a novel framework based on Clustering Analysis for the definition of driving simulator experiments [2]. Amir Mohammadi Amiri et al. (2021) used five different hotspot identification algorithms. They are as follows: Getis-Ord Gi*, Average Nearest Neighbor, kernel density KDE, Global Moran's I, and mean center. Global Moran's I approach outperforms other methods in locating hotspots, according to the findings [3]. In the same year, F. Francis used Hierarchical clustering and K-means clustering the same year to merge the spatially specified groupings into six clusters based on the similarity of their temporal patterns [4]. Dooti Roy et al. (2021) introduced a two-stage clustering-based technique based on SOM followed by neural gas clustering to build a data-driven taxonomy of bus crashes [5]. Rocio Suarez-del Fueyo et al. (2021) used unsupervised clustering methods to identify badly injured, belted occupants into groups, bio-mechanical characteristics, and accident severity [6]. The applicability of the k-prototypes clustering method in massive truck-involved crashes was investigated by Syed As-Sadeq Tahfim et al. (2021). To predict the severity of injuries in major truck incidents, four gradients boosted decision trees techniques were used to the dataset and individual clusters [7]. Filbert Francis et al. (2021) found high-risk areas in Dar es Salaam for motorcycle-related injuries. Three distinct motorcycle injury hotspot clusters have been discovered [8].

Mert Ersen et al. (2020) used the Kernel Density approach to examine statistical analyses based on accident kinds. The

Kernel Density approach has been found to produce better visual results than other spatial methods [9]. Seyed Mohsen Hosseinian et al. (2020) investigated the effect of different factors on the severity of urban traffic accidents in Rasht metropolis by using frequency analysis of accident data [10]. Qiuru Cai (2020) created the Apriori algorithm to mine the rules that govern the relationship between risk issues and the cause of traffic accidents on urban roads [11].In 2020, Yunduan Lin et al. used crowdsourcing data to investigate the technique of predicting the complicated behavior of traffic flow evolution after traffic accidents. According to the results, NN outperforms the other models [12].

Sharaf AlKheder et al. (2020), on the other hand, used three data mining algorithms to conduct a thorough investigation of risk factors associated to the severity of traffic accidents. In comparison to previous models, the Bayesian network was more accurate in predicting the variables [13]. Yang Yong Zheng et al. (2020) discovered the elements that influence traffic accidents in undersea tunnels and developed a prediction model for undersea tunnel traffic accidents [14]. Marjana Cubranic-Dobrodolac et al. (2020) suggested a model for assessing and making decisions about a driver's proclivity for traffic accidents that is based on an estimation of the driver's psychological attributes [15].

Based on single-vehicle crashes, Natalia Casado-Sanz et al. (2020) found the contributing factors to a fatal outcome. The most relevant factors related with driver injury severity were identified using a Multinomial Logit model [16]. Human error was highlighted as a major contributory element in road traffic accidents by Asad Iqbal et al. (2020), and the Salt Range was classified as a black spot-on account of vehicle braking failure [17]. Minglei Song et al. suggested a road accident prediction model based on joint probability density feature extraction from big data in 2019 [18]. Eight impact factors were chosen by Cheng Zhang et al. (2019), and the Bayesian network was the best model to potentially predict road accident black spots [19].

On accident datasets, Sadiq Hussain et al. (2019) used J48, Multi-layer Perceptron, and BayesNet classifiers. The Multi-layer Perceptron classifier per-formed well in the study, with an accuracy of 85.33 percent [20]. According to Juan Pineda-Jaramillo et al. (2019), road traffic collisions occur in all clusters, although zones surrounded by landscapes and parks have more run overs than fallen residents [21].

III. Materials and Methods

A. Dataset Description

From 2007 to 2015, we collected 1283 data of traffic incidents on the N5NH (N5 National Highway) from MAAP5 of the Accident Research Institute (ARI), BUET Accident report forms have been distributed to various police stations in Bangladesh. There are two parts to the accident report form. One is the main form, while the other is the supplementary form. Each accident record is filled out on the main form, where the top 37 columns depict preliminary information on the severity of the traffic accident. Accidental vehicle information is stored in columns 38-45, whereas driver information is stored in columns 68-52. Furthermore, columns 53-58 and 59-64 contain detailed information about passengers and pedestrians, respectively. 65-67 columns, on the other hand, are utilized to identify the causes of an accident.

B. Proposed Discriminatory Factors of Fatal Subtype Detection Model

Fig. 1 depicts an overview of our process for identifying discriminatory characteristics, which is briefly detailed step by step.

Step 1: Data Preprocessing and Analysis: In this section, all portions of the data, including the route number, have a recurring value. Furthermore, some data, such as XY map, X coordinate, and Y coordinate, have no values. 67 percent values are also missing in the kilometer post and 100-meter attributes. As a result, we decide experimental data eliminate them. We have separated 1002 fatal data from 1283 entries. Numeric and nominal values are blended throughout all records. All nominal values have been converted to numeric values. The features dealing with vehicle details (columns 38-45), driver details (columns 46-52), passenger details (columns 53-58), and pedestrian details (columns 53-58) are deleted from empirical traffic accident data as unusable features. Table I illustrates these characteristics with a brief explanation. Report Number, FIR Number, and Thana are not deemed particularly important and are detached from the empirical data. Formerly, we construct a hit-map to detect linked traits (see Fig. 2). So, we see that the number of vehicles is correlated to the number of driver and pedestrian victims. As a result, we eliminate these two attributes. The remaining attributes are useful in determining the more accurate outcomes in this experiment.

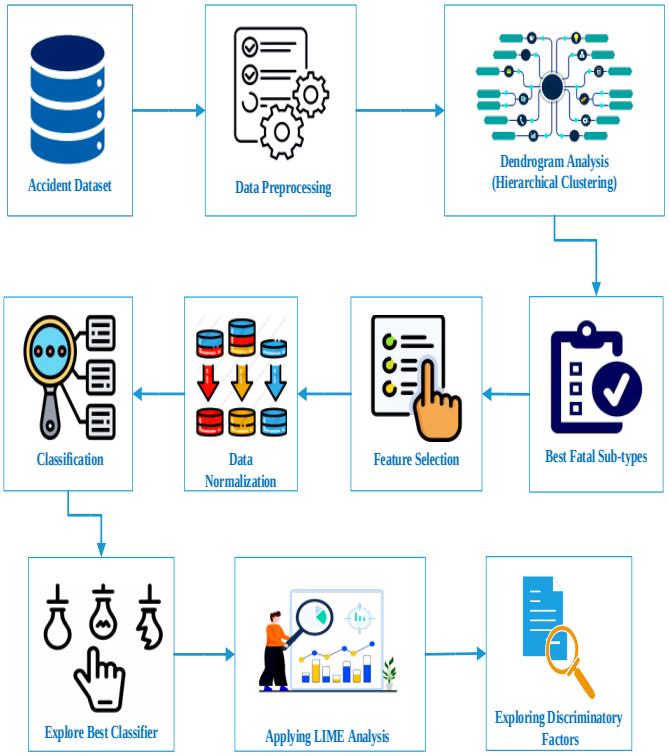


Fig. 1. Proposed Identifying Discriminatory Factors of Fatal Subtype Model.

TABLE I. ACCIDENT DATASET DESCRIPTION

Sl.	Feature Name	Feature Description
1.	Report Number	Accident report number.
2.	FIR Number	Accident FIR number.
3.	Thana	Thana number.
4.	District	District name.
5.	Number of Pedestrian Victims	Pedestrian victims' number.
6.	Number of Passenger Victims	Passenger victims' number.
7.	Number of Driver Victims	Driver victims' number.
8.	Number of Vehicles	Vehicles intricate numbers.
9.	Day of Week	The casualty occurred the day.
10.	Month	The casualty occurred a month.
11.	Date of Month	Casualty date.
12.	Year	The casualty occurred a year.
13.	Accident Type	Collision type of occurred accident.
14.	Movement	Road variety.
15.	Type of Junction	Diversity of junction.
16.	Traffic Control	The behavior of regulating traffic.
17.	Divider	Existents of the divider.
18.	Weather	Condition of weather when the accident occurred.
19.	Light	Light condition on the road surface.
20.	Road Geometry	The geometry of road surface.
21.	Severity	Type of a casualty.
22.	Type of Surface	Variation of the road surface.
23.	Condition of Surface	Condition of the road surface.
24.	Surface Quality	The road quality.
25.	Type of Location	Accident location.
26.	Read Feature	Variation of road.
27.	Road Class	Road Category.
28.	Responsible Factors 1	An accident-related factor.
29.	Responsible Factors 2	An accident-related factor.
30.	Responsible Factors 3	An accident-related factor.

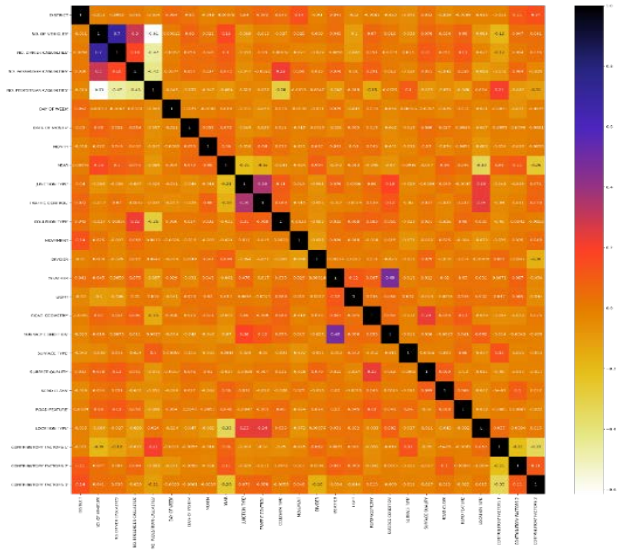


Fig. 2. Hit-Map for Identifying Correlated Features.

Step 2: Employing Clustering Analysis: Clustering analysis is a technique for categorizing cases into comparable important groups based on their unique characteristics. The agglomerative mode of hierarchical clustering algorithms divides every cluster into small sub clusters or assembles them into super clusters on a regular basis [2]. In a hierarchical architecture known as a dendrogram, the connection between each pair of clusters is determined by the medium of dissimilarity or similarity. However, we apply this strategy to generate numerous fatal sub-types in the accident dataset. To reveal the predictability of the proposed model, these subtypes are considered as separate class labels.

Step 3: Chi-Square Test for Feature Ranking: When two attributes are independent, the executed count is close to the awaited count, resulting in a reduced Chi-Square value. The higher the Chi-Square number, the more dependent the property is on the response. Then it can be chosen for model training. However, in our research, we rank attributes in order to identify the appeasement set of most significant factors that result in the maximum accuracy. After identifying the sub-types, we utilized the Chi-Square test feature ranking technique on the accidental dataset to discover the optimal set of most significant attributes.

Step 4: Normalization: Normalization is a data preparation method used frequently in machine learning. Its major purpose is to use a common scale to adjust the values of numeric columns in the dataset without losing information. In this paper, we use the MinMaxScaler class in Python to normalize fatal sub-types data of the most significant attributes and create a balanced dataset with appropriate structures.

Step 5: Classification Approach: Classification Approach: To compute the class of objects, Classification is a mode of function discovery in which concepts or classes are interpreted and isolated whose label is unfamiliar to the target. On the normalized dataset, we use six machine learning classification algorithms to identify the observed sub-types: Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Multi-layer Perceptron (MLP and Support Vector Machine (SVM). Previous studies of road accidents have used these classifiers extensively [12, 19]. To find the best classifier with the highest accuracy, some evaluation metrics (see Table II) such as Accuracy, F1-Score, and AUROC were used.

Step 6: Exploring Discriminatory Factors Using LIME: Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm that can interpret a model by distracting the data sample input and knowing how the predictions vary. LIME produces a set of interpretations that show how each feature performs against a prediction for a single sample, which is a type of local interpretability. We use LIME to find which characteristics contributed the most to attaining the best result in categorizing the sub-types on the dataset using all the attributes for the best per-forming classifier. As a result, we have discriminatory variables for the classification of subtypes.

TABLE II. EVALUATION MATRIX

Metrics	Details	Formula
Accuracy	Accuracy is the ratio of the number of all correct predictions and the total number of the data.	$Acc. = \frac{TP+TN}{TP+FN+FP+TN}$
F1-Score	F1-Score is a harmonic mean of precision and recall.	$F = \frac{2 * Recall * Precision}{Precision + Recall}$
AUROC	AUC is the summation of all TP rates and TN rates divided by two.	$AUROC = \frac{TP\ rate + TN\ rate}{2}$

IV. RESULT AND DISCUSSION

In our work, we identified the clusters using hierarchical clustering on the dataset. Each cluster is defined as an observed subtype present in the accident dataset. We utilized the Chi-Square test to determine the most significant features after identifying the subtypes. Then, on the selected features, we performed data normalization using Python's MinMaxScaler class. On the datasets, we used various classification algorithms (i.e., DT, KNN, NB, RF, SVM, MLP) to classify the observed subtypes. Classification is accomplished through the use of 10-fold cross-validation. Finally, we used LIME to interpret features for discriminatory factors. Jupyter Notebook version 6.1.4 is used for all of the experiments.

A. The Analogy of Performance of Distinct Classifiers

In this study, Hierarchical clustering yields two subtypes (subtype-1 and subtype-2) (see Fig. 3). The ratio of subtype-1 to subtype-2 is found to be the same. As a result, no data balancing was required. The Chi-Square test result is displayed in Table III. As can be seen in the table, the features are ordered in ascending order depending on their P-Values. The feature with the lower P-Value is more important. We took the different number of features (i.e., 5, 10, 15, 20, and 24) from those significant feature lists and applied different classifiers to the datasets. The experimental results for different classifiers (described in section *III.B*) utilized to categorize the sub-types are shown in Tables IV, V, VI, VII, and VIII. To explain our findings, we used a variety of evaluation matrices (Accuracy, F1- score, and AUROC). Performance Analysis of All Significant Features is shown in Fig. 4.

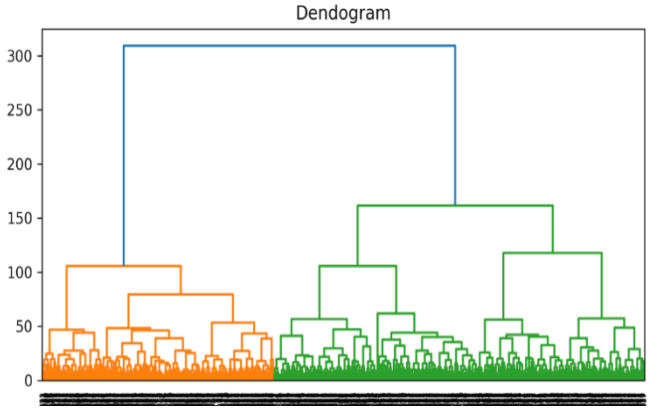


Fig. 3. Dendrogram Analysis to Identify the Fatal Subtypes.

TABLE III. P-VALUES IN ASCENDING ORDERS FOR DIFFERENT FEATURES

Features	P-Value
Date of Month	0.000000
Number of Passenger Victims	0.000017
Month	0.000634
District	0.003480
Surface Quality	0.010259
Type of Junction	0.015621
Light	0.047291
Accident Type	0.059001
Condition of Surface	0.240108
Type of Location	0.252266
Responsible Factors 1	0.419230
Number of Vehicles	0.479777
Day of Week	0.495156
Traffic Control	0.510946
Weather	0.545987
Divider	0.582690
Responsible Factors 2	0.681025
Road Class	0.885795
Year	0.895612
Movement	0.942830
Responsible Factors 3	0.954226
Road Geometry	0.957346
Road Feature	0.988516

TABLE IV. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (5 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.80	99.80	99.80
KNN	91.91	91.91	97.91
NB	97.00	97.00	98.81
RF	**99.90**	**99.90**	**100.00**
SVM	97.60	97.60	99.77
MLP	99.30	99.50	99.94

TABLE V. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (10 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	**99.80**	99.60	99.80
KNN	77.34	77.34	85.07
NB	95.61	95.61	98.10
RF	**99.80**	**99.80**	**100.00**
SVM	95.21	95.21	99.11
MLP	98.20	98.60	99.78

TABLE VI. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (15 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.60	**99.70**	99.50
KNN	65.06	65.06	70.74
NB	94.91	94.91	97.96
RF	**99.70**	**99.70**	**100.00**
SVM	92.32	92.32	98.29
MLP	97.60	97.70	99.76

TABLE VII. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (20 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.50	**99.60**	99.50
KNN	62.26	62.26	65.39
NB	93.61	93.61	97.31
RF	**99.70**	**99.60**	**100.00**
SVM	92.31	92.31	98.10
MLP	97.10	96.70	99.65

TABLE VIII. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (24 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.60	99.60	99.50
KNN	55.38	55.38	57.79
NB	92.71	92.71	97.05
RF	**99.80**	**99.90**	**100.00**
SVM	90.31	90.31	97.62
MLP	96.31	96.31	99.61

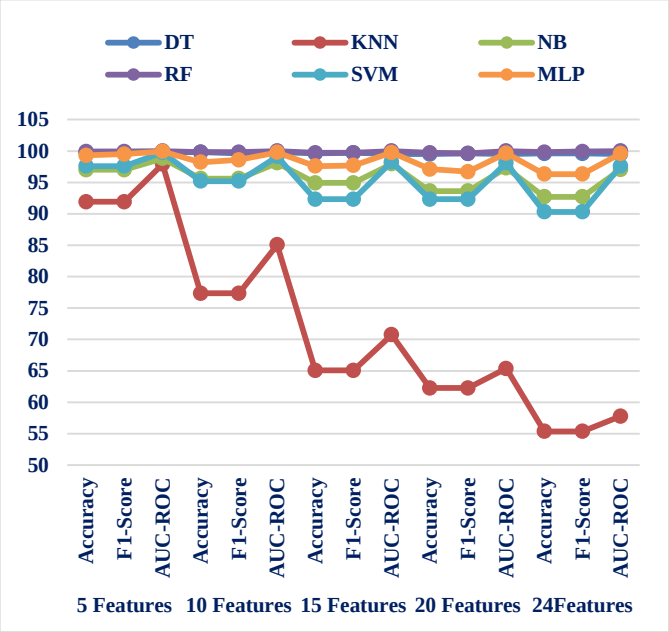


Fig. 4. Performance Analysis of All Significant Features.

From the table, we can see that RF outperforms all other classifiers in terms of accuracy, F1-score, and AUROC (see Table IV to VIII) for all different number of features. For the dataset with only the five most significant features, RF achieves 99.90% accuracy, 99.90% F1-score, and 100.00% AUROC. This is the highest possible score in our study. For other sets of features, RF receives slightly different scores. Furthermore, all other classifiers, with the exception of KNN, achieve high results (i.e., above 90%). It is also worth noting that all classifiers performed best with the most significant 5 feature subset, and their performance degraded as the number of features used increased.

B. Interpretation of Features for Discriminatory Factors

We used LIME on the dataset (with all features) to find the highest performing RF classifier and determine which features contributed the most to correctly categorizing the subtypes. As a result, we obtain discriminatory factors for sub-type classification. The features that contributed the most to identifying distinct sub-types are shown in Fig. 5, which differs significantly from the statistical result we obtained using the Chi-Square test for important features. The most crucial feature identified for subtype classification is 'Road Feature,' as seen in Fig. 5. The relationship between road features and accident subtype is seen in Table IX. The table shows that "Road Feature" - General is the most prevalent cause of accidents and has about the same ratio in both categories. "Road Feature"- Bridge is twice as common in subtype-1 as in subtype-2. Culverts and Speed Breakers are more common in subtypes 1 and 2, respectively. The second most significant attribute is 'Road Class.' The relationship between road classes and accident subtypes is shown in Table X. It is apparent that the most prevalent type of accident is 'Road Class'- Natural and has nearly the same ratio in both subtypes. 'Road Class'- Feeder' is twice as common in subtype-1 as it is in subtype-2. 'No. of Vehicles' is the third most essential aspect. The relationship between the 'No. of Vehicles' and the types of accidents is seen in Table XI.

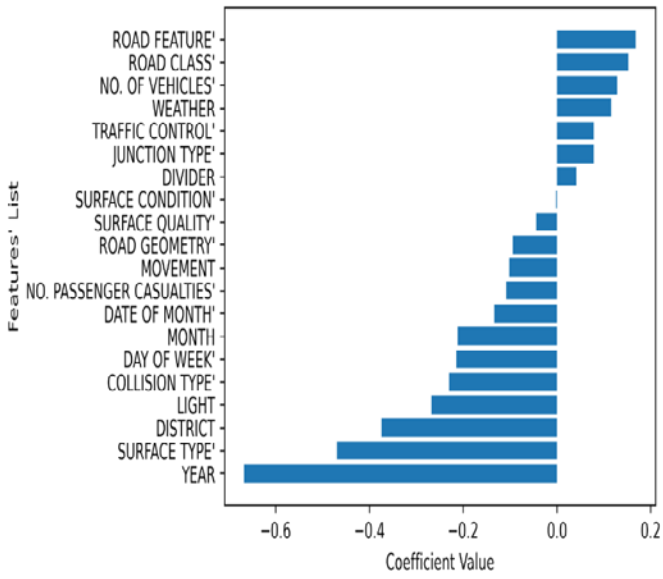


Fig. 5. Interpreting Important Features using LIME.

TABLE IX. THE RELATION BETWEEN "ROAD FEATURE" AND ACCIDENT SUBTYPE

Road Feature	Cluster	No. of instances
General	Subtype-1	484
	Subtype-2	486
Bridge	Subtype-1	14
	Subtype-2	7
Culvert	Subtype-1	1
	Subtype-2	3
Speed Braker	Subtype-1	2
	Subtype-2	3
Narrow	Subtype-1	1
	Subtype-2	1

TABLE X. THE RELATION BETWEEN "ROAD CLASS" AND ACCIDENT SUBTYPE

Road Class	Cluster	No. of instances
Natural	Subtype-1	488
	Subtype-2	486
Feeder	Subtype-1	4
	Subtype-2	2
Regional	Subtype-1	6
	Subtype-2	8
Rural	Subtype-1	2
	Subtype-2	3
City	Subtype-1	2
	Subtype-2	1

TABLE XI. THE RELATION BETWEEN "NO. OF VEHICLES" AND ACCIDENT SUBTYPE

No. of Vehicles	Cluster	No. of instances
1	Subtype-1	299
	Subtype-2	265
2	Subtype-1	198
	Subtype-2	234
3	Subtype-1	4
	Subtype-2	1
5	Subtype-1	1

TABLE XII. THE RELATION BETWEEN "WEATHER" AND ACCIDENT SUBTYPE

Weather	Cluster	No. of instances
Clean/Fair	Subtype-1	461
	Subtype-2	458
Fog	Subtype-1	18
	Subtype-2	25
Rain	Subtype-1	21
	Subtype-2	17
Wind	Subtype-1	2

The table shows that for subtype-1, 'No. of Vehicles' 1 is more prevalent, while for subtype-2, 'No. of Vehicles' 2 is more common. 'No. of Vehicles' 5 appears only in subtype-1. 'Weather' is the fourth most essential feature. The relationship between weather and accident subtypes is seen in Table XII. The data shows that 'Weather'- Clean/ Fair has a higher risk of accidents. 'Weather'-Rain is exclusively related to the subtype-1. As shown in Fig. 5, 'Traffic Control,' 'Junction Type,' and 'Divider' all play a role in subtype classification. Other features in the list have a negative relationship with classification into subtypes.

C. Relative Studies and Implication

Many researchers have looked into road accident classification, and some of their findings are included in Section 2. We discovered that all of the research efforts on these datasets were focused on classifying the risk of a traffic collision. However, no attempt was made to identify the various forms of road accidents (as far as our knowledge). We used clustering to determine the subtypes in this study, and the appropriate number of clusters for each dataset is justified. Then we used classifiers to find the best classification of subtypes using relevant feature sets. Then, using the explainable AI technique, we showed key features that contributed to the identification of subtypes. Identifying subtypes will assist authorities in better understanding accident risks. We discovered important elements that will assist them in identifying sub-types as well as accident risks.

V. CONCLUSION AND FUTURE WORK

Traffic accidents are viewed as a global issue that results in fatalities and serious injuries. The study of traffic accident data assists the traffic department in identifying the primary persuasive elements of accidents and revealing the relationships between these issues, creating the groundwork for risk control measures to be developed. Discriminatory factors can increase the likelihood of traffic accidents or other factors that contribute to the severity of injuries sustained as a result of traffic accidents. We've compiled a list of 24 features that are thought to be linked to road accidents. The information was obtained from the Accident Research Center (ARI) at BUET. To recognize the clusters, we utilized hierarchical clustering on the dataset. Each cluster represents a perceived subtype found in the accident dataset. To categorize those experimental subtypes, we applied six different classification methods on the datasets. Finally, for the interpretation of features for discriminatory variables, we used the LIME analysis technique. As a result, future work will necessitate a thorough examination of the updated dataset of traffic accidents from across the country, as well as the application of more classification and clustering algorithms, as well as the improvement of the discriminatory factors identification model through additional development and experiments to conduct follow-up traffic accidents.

ACKNOWLEDGMENT

We are grateful to the Accident Research Center (ARI), BUET (http://ari.buet.ac.bd/) for allowing us to collect information and data from its Modular Accident Analysis Program, Version-5 database software.

REFERENCES

[1] N. F. Antara, "Dhaka Tribune," E-News Paper, 9 1 2021. [Online]. Available:https://www.dhakatribune.com/bangladesh/2021/01/09/report-18-people-killed-every-day-on-average-in-road-accidents-in-2020. [Accessed 20 10 2021].

[2] Bobermin, Mariane, and Sara Ferreira. "A novel approach to set driving simulator experiments based on traffic crash data." Accident Analysis & Prevention 150 (2021): 105938.

[3] Amiri, Amir Mohammadian, et al. "GIS-based crash hotspot identification: a comparison among mapping clusters and spatial analysis techniques." International journal of injury control and safety promotion (2021): 1-14.

[4] Li, Yu, et al. "Crash report data analysis for creating scenario-wise, spatio-temporal attention guidance to support computer vision-based perception of fatal crash risks." Accident Analysis & Prevention 151 (2021): 105962.

[5] Roy, Dooti, Ved Deshpande, and M. Henry Linder. "A cluster-based taxonomy of bus crashes in the United States." Computational Statistics (2021): 1-18.

[6] Suarez-del Fueyo, Rocio, et al. "Cluster analysis of seriously injured occupants in motor vehicle crashes." Accident Analysis & Prevention 151 (2021): 105787.

[7] Tahfim, Syed As-Sadeq, and Chen Yan. "Analysis of Severe Injuries in Crashes Involving Large Trucks Using K-Prototypes Clustering-Based GBDT Model." Safety 7.2 (2021): 32.

[8] Francis, Filbert, et al. "Investigation of road infrastructure and traffic density attributes at high-risk locations for motorcycle-related injuries using multiple correspondence and cluster analysis in urban Tanzania." International journal of injury control and safety promotion (2021): 1-11.

[9] Ersen, Mert, Ali Hakan Büyüklü, and Semra Erpolat Taşabat. "Analysis of Fatal and Injury Traffic Accidents in Istanbul Sarıyer District with Spatial Statistics Methods." Sustainability 13.19 (2021): 11039.

[10] Hosseinian, Seyed Mohsen, and Vahid Najafi Moghaddam Gilani. "Analysis of factors affecting urban road accidents in rasht metropolis." Eng Transactions 1 (2020): 1-4.

[11] Cai, Qiuru. "Cause Analysis of Traffic Accidents on Urban Roads Based on an Improved Association Rule Mining Algorithm." IEEE Access 8 (2020): 75607-75615.

[12] Lin, Yunduan, and Ruimin Li. "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data." Accident Analysis & Prevention 145 (2020): 105696.

[13] AlKheder, Sharaf, Fahad AlRukaibi, and Ahmad Aiash. "Risk analysis of traffic accidents' severities: An application of three data mining models." ISA transactions 106 (2020): 213-220.

[14] Yong-Zheng, Yang, and Y. Mei. "Analysis of Influencing Factors of Traffic Accidents in Undersea Tunnel." Journal of Engineering Research and Reports (2020): 37-49.

[15] Čubranić-Dobrodolac, Marjana, et al. "Modelling driver propensity for traffic accidents: a comparison of multiple regression analysis and fuzzy approach." International journal of injury control and safety promotion 27.2 (2020): 156-167.

[16] Casado-Sanz, Natalia, Begoña Guirao, and Maria Attard. "Analysis of the risk factors affecting the severity of traffic accidents on Spanish crosstown roads: the driver's perspective." Sustainability 12.6 (2020): 2237.

[17] Iqbal, Asad, et al. "Road Traffic Accident Analysis and Identification of Black Spot Locations on Highway." Civil Engineering Journal 6.12 (2020): 2448-2456.

[18] Song, Minglei, Rongrong Li, and Binghua Wu. "A novel prediction model of traffic accidents based on big data." International Journal of Modeling, Simulation, and Scientific Computing 10.04 (2019): 1950022.

[19] Zhang, Cheng, Yue Shu, and Lixin Yan. "A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China." IEEE Access 7 (2019): 140197-140205.

[20] Hussain, Sadiq, et al. "Performance evaluation of various data mining algorithms on road traffic accident dataset." Information and Communication Technology for Intelligent Systems. Springer, Singapore, 2019. 67-78.

[21] Pineda-Jaramillo, Juan, and Óscar Arbeláez-Arenas. "Modelling road traffic collisions using clustered zones based on Foursquare data in Medellín." Case Studies on Transport Policy 9.2 (2021): 958-964.

A Review on Software Bug Localization Techniques using a Motivational Example

Amr Mansour Mohsen[1], Hesham Hassan[2], Ramadan Moawad[3], Soha Makady[4]
Computer Science Department, Faculty of Computers and Information Technology, Future University in Egypt, Cairo, Egypt[1, 3]
Computer Science Department, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt[2, 4]

***Abstract*—Software bug localization is an essential step within the software maintenance activity, consuming about 70% of the time and cost of the software development life cycle. Therefore, the need to enhance the automation process of software bug localization is important. This paper surveys various software bug localization techniques. Furthermore, a running motivational example is utilized throughout the paper. Such motivational example illustrates the surveyed bug localization techniques, while highlighting their pros and cons. The motivational example utilizes different software artifacts that get created throughout the software development lifecycle, and sheds light on those software artifacts that remain poorly utilized within existing bug localization techniques, regardless of the rich wealth of knowledge embedded within them. This research thus presents guidance on what artifacts should future bug localization techniques focus, to enhance the accuracy of bug localization, and speedup the software maintenance process.**

Keywords*—Bug localization; bug localization artifacts; information retrieval; program spectrum

I. Introduction

Software maintenance is considered a continuous process in software projects. However, software maintenance is one of the most expensive stages in the software development life cycle [1]. According to Erlikh [2], maintenance consumes 70% and maybe up to 90% of the time of any product's life cycle. In addition to that, Hunt et al. [3] presented that the maintenance process takes above 50% of the software life cycle. Also, Lientz and Swanson [4] claimed that software maintenance spending from 20% to 70% of the efforts exerting on maintenance. Software Maintenance is defined by Sommerville [5] as "the modification of a software product after delivery to correct faults, to improve performance or other attributes". Software maintenance must be applied to improve the design, implement enhancements, and interface with other legacy software [6] to build a new one with some updates or solve bugs.

A different view of software maintenance [7] defines it as "error, flaw, or fault in a computer program or system that produces unexpected results or behavior". Once the bug occurs, the bug triaging and localization process is applied to solve the bug [7]. The process involves: (i) understanding the bug, (ii) assigning a maintainer, and (iii) bug localization within the source code, and (iv) bug fixing. The bug localization process is the action of determining the location of the bug in the software program [8]. However, locating the bug manually could be time consuming, cost consuming, and infeasible [10].

Several techniques have been utilized to localize bugs automatically, including: information retrieval [9], machine learning. [10], program spectrum [11], and program slicing [12]. Those techniques use different software artifacts like bug reports, stack traces, source code files. However, such techniques do not necessarily benefit from all the information present within those artifacts. For instance, techniques that utilize source code do not use the structural relationships between source code elements to locate bugs, although such information could improve the accuracy of bug localization. Hence, a review is conducted to identify the different artifacts utilized by bug localization techniques, and how well such artifacts' information gets utilized.

Furthermore, a motivational example is introduced. Within such motivational example, we present a running example that includes different software artifacts and a set of injected bugs. We applied various existing bug localization techniques that utilized subsets of the included artifacts, to locate the injected bugs within such example, and assessed various bug localization techniques on those bugs. What difference in this review that the process of motivational example helps in identifying the limitations of those bug localization. Besides it gives perspective for better utilization of the different software artifacts to increase the quality of the results of such bug localization techniques.

The rest of this paper is structured as follows. Section II will present the related works to software bug localization techniques. Section III presents the motivational example and its software artifacts. Three categories of bug localization techniques: information retrieval, machine learning, and program spectrum will be explained, and applied to the motivation example within Sections IV, V, and VI respectively. Section VII discusses the findings and concludes the review.

II. Related Work

A. Related Work on Information Retrieval

An information retrieval technique called (BLIA) bug localization using integrated analysis [9] proposed by Klaus Changsun et al. to illustrate the technique besides showing limitations. Such work utilizes different software artifacts like stack traces, comments, bug reports, and the history of code modifications are features utilized in the work.

Klaus Changsun et al. evaluated their work on three open-source projects: Aspect-oriented extension to Java (AspectJ), Widget toolkit for Java (SWT), and Barcode image processing

library (Zxing). The number of bugs and source files that they worked on are as follows: AspectJ (284 bugs, 5188 source files), SWT (98 bugs, 738 source files), and Zxing (20 bugs, 391 source files). Five steps were followed to complete their approach. First, an information retrieval technique is used for measuring the similarity between the bug reports text and source code files called rVSM [13]. Then the structured data in the bug reports like bug description, bug summary and other stated before in the bug report artifact are analyzed and integrated with the above data. After that, if stack traces appeared in the bug report, they would be analyzed to extract the beneficial information to improve the results of retrieval. Moreover, the historical data for code modifications which is extracted from version control systems to predict the affected files and methods. Finally, similarity measurements were applied between the accumulative data from the above steps and the source code files. The results will be ranked by scores to the files of the code which is mainly expected to have the error. The proposed approach resulted in an enhancement in the mean precision over some other approaches like BugLocator with 54%, BLUiR with 42%, and BRTracer with 30%, and Amalgam with 25%.

Wen et al. proposed an information retrieval technique to localize bugs called FineLocator. FineLocator recommends the position of bugs based on method level [14]. It means that not only recommend the source file that contains bug but also the method contains bug. The proposed architecture consists of three main components are method extraction, method expansion, and method retrieval. The method extraction process is applied by extracting the methods names and their bodies using the abstract syntax tree for the code. Additionally, the timestamp for the methods is also extracted from version history systems and the dependence information for each method is also extracted. The first sub-component of method expansion is the semantic similarity measurement between methods. This step will be applied first by generating a numeric vector for each method by generating a bag of words and among all methods of the code. Then the scores are calculated between every two methods to know the similarity score between them. Then the call dependency is applied among the class level and the method level to enhance the similarity scores between methods. Besides, another score is calculated which is temporal proximity measure which calculates the difference in time of edit between the methods as the methods that edited in time near each other will be more probable to be near to each other. Then all the above scores are combined to one value which is the method augmentation value. They test their work on ArgoUML, Maven, Kylin, Ant, and AspectJ and enhance the performance of the method level by 20% MRR.

Yaojing et al. [15] proposed an approach with three main considerations which are 1) the fix history relationships with old bug reports, 2) word co-occurrence in the bug reports and source files, 3) The long source files. The proposed model consists of a supervised topic modeling technique called LDA for classifying the old bug reports and bug reports with special topic w. Then the word co-occurrence with words from bug reports that appear in the bug reports. In addition to the creation of the long source files and stack traces in bug reports. They test their work on 10-fold cross-validation. Also, the proposed model was applied to three main projects PDE with 3900 bug reports and 2319 source files, the platform with 3954 bug reports and 3696 source files, and JDT with 6267 bug reports and 7153 source files.

Mills et al. [16] constructed an approach trying to enhance the process of text retrieval bug localization by studying the most important elements of a bug report. A genetic algorithm is applied to find the optimal query to retrieve the true results from source files. Yu Zhou et al. construct an approach [17] that consists of three steps to classify bug reports: Classifying the summary part of each bug into (high, middle, and low) using a machine learner. It will help to increase the accuracy of bug localization systems. Then some structured features are used from the bug reports using a machine learner.

Additionally, the results are merged from the above steps and other machine learning algorithms are used. The authors manually classify the bug reports into six categories (BUG, RFE, IMPR, DOC, REFAC, other). Additionally, a voting is applied [18] between different developers to classify each bug report to label them. They need to classify either the bug report is a bug or not. They answer the question of that a given report is a corrective bug or not by using different fields in the bug report. Also, the proposed approach Combines text mining and a data mining approach to solve the problem. The approach evaluated using 3200 random reports from large projects like Mozilla, Eclipse, JBoss, firefox, and OpenFOAM. The Use Bugzilla as the bug tracking system. They use the only reports that are tagged by resolved or closed to analyze them. They consider multiple fields of the bug report like (textual summary, severity, priority, component, assignee, and reporter.

Alessandro Murgia et al… Tonelli [19] tried to make bug tracking systems linked with CVS to enhance the bug fixing and relations between different versions of the software and the bugs and also the end-users. Each commit component consists of (author when it was done, modified files, and commit messages). The work was stressed on fixing-issue commits. They manually labeled the data of commits to training their classifier through one author and this is a drawback as the author may do not know enough the data in the commits then maybe the classifier is biased to their labeling. Preprocessing steps from natural language processing are used like stemming and stop words removal to enhance the classifier. Additionally, some regular expressions are used to filter commits that relate to specific bugs. The features used to feed the classifier are the words extracted from the commits. They applied their experiments to Netbeans and Eclipse projects. The machine learning classifier got a precision of 99.9% for classifying fix issue and non-fix issue commits. The dataset used has not appeared as they didn't use a benchmark dataset. Besides, they identify the main terms used for bug-fixing issues like the fix, for, and bug. The support vector machines are classified with accuracy up to 99.9%.

B. Related Work on Machine Learning

In [52], the authors produced an approach for localizing the bugs automatically using ranking. The source code files are

ranked to the most probably that contains the bug reported. Different features will be used as a bag of words used from source code files and bug reports. The similarity that is measured between bug reports and source code files using cosine similarity. Also, the API information is used to enhance the features. Another feature is collaborative filtering which is applied between similar bug reports. Additionally, the class names and the bug fixing frequency considered to be featured. They apply their experiment on AspectJ, Eclipse UI, JDT, SWT, and Tomcat. The average accuracy of 70% achieved all over the top 10 ranked files.

In [20], an approach proposed using deep learning with rVSM to enhance the process of bug localization. The revised vector space model (rVSM) is utilized to set up the features that are used in measuring the similarity between bug documents and source code files. The DNN is used to measure the relevancy of the term between the terms in bug reports and source code files. Also, another type of feature rather than terms is the metadata feature about source code files, it seems like logs about the file. The inputs are text similarity, metadata about source code files. They used DNN to learn all the features. They applied on different datasets like AspectJ, Birt, Eclipse UI, JDT, SWT, and Tomcat. They got an average precision of 0.52 using the tomcat dataset.

Dongsun Kim, Sunghun Kim, and Andreas Zeller proposed a model [21] with two phases to predict the files to be fixed. The bug report in many cases as mentioned by the authors may not contain sufficient information to help in predicting the files needed to be fixed. A machine learning approach is applied to classify the bug reports as predictable which means contain useful information or not predictable. The Features extracted from the bug reports are the summary, platform, operating system, severity, priority, and reporter. Then the model is trained using the specified machine learning and tested. Then in phase two, the predictable bug reports to be fixed are then entering a multi-class classification model to know the exact files to be fixed. The recommended model was evaluated using 70 percent of the dataset for training and 30 percent for testing. They achieved an average accuracy for predicting files to be fixed with 70 percent.

ERIC et al. [22] proposed a neural networks technique based on the code coverage data as a feature. This coverage data comes from applying virtual test cases to each line in the code. Then they feed them to a neural network. The technique was tested on four different benchmark datasets (Siemens, UNIX, Grep, and Gzip). They enhance the performance of examining lines of code than [23].

In [24], [25] a deep learning model are applied in order to localize bugs using source code files and bug reports. They got accuracy of applying on different benchmark datasets.

Liang et al. [10], proposed a deep learning system to localize bugs. Bug reports text terms are utilized besides the terms of source code files. The works are evaluated on four datasets (AspectJ, SWT, JDT, and Tomcat) with the following MAP (0.439, 0.457, 0.482, and 0.561).

C. Related Work on Program Spectrum

Jeongho et al. proposed a spectrum-based technique that localizes bugs based on the variables that are most probably suspicious [11] to rank the lines most probably contain bugs. A limitation discussed in this paper about previous work considering program spectrum that if there is an else block as an example and the block contains many lines. The outcome of the ranking of lines contains code will not be accurate and maybe the cause of the error be directly before the block. To overcome the above limitation, the variable-based technique proposed to keep track of mainly the information about the suspicious variables and their coverage in the code. First, the variable spectra are created by using the test cases as an input in addition to the execution trace data for each variable. Then the suspicious ratios are calculated by substituting the variable spectra with the coefficient's similarity. The final step is applied by rank the most variables that are most suspicious in descending order to the bug solver. The work was evaluated using the Exam score evaluation metric.

On the other side, Henrique et al. constructed a spectrum-based fault localization tool called Jaguar which stands for Java coverage fault localization ranking [26]. An architecture was formulated for the tool consists of two main components which are Jaguar Runner and Jaguar Viewer. The java runner component gathers the data for control flow spectra and data the data flow using different unit tests. After the data collection steps applied, then a metric score calculated using one of past known calculations Metric like [27]. After that, the mixed scores between data and control flow matrices are normalized for the suspicious parts of the code. Then the jaguar viewer colors the suspicious entities of the code according to their score with for different colors according to their danger. They assessed t their work based on the Defects4J dataset.

A new method that depends on the level of predicates not all the lines of the code was constructed by B´ela that utilizing the data from test cases and code coverage data [28]. This special type of spectrum-based fault localization took into consideration which methods will be hit in the run time of test cases to use these data in ranking the most suspicious methods. Additionally, different past research metrics for ranking that used for the lines of code as stated in [29] will be used at the method level. The pre-step to the algorithm is the building of the coverage matrix between the methods and the test cases. A graph will be generated from the coverage as the nodes of the graph represent the methods and the tests. The edges that will link different nodes with each other represent that a node that may be a test case will hit a node which is a method. Besides, the failed test cases will be marked in the graph. The first step was to calculate the edge weights by summing up the total methods that hit a failed test case to all methods. Then the values will be updated by calculating the average value of methods that cover failed test cases. The next step is to aggregate the values of edges to the method nodes. Finally, the nodes of methods values will be updated by calculating the resulted values concerning the number of test cases. They evaluated their work based on the Defects4J dataset that includes four projects with good results of the ranking.

Abubakar et al. proposed a graph-based technique for the spectrum-based technique based on the execution of the test cases [30]. The technique aims at localizing not only a single bug in the system but also multiple bugs during execution. The exploration of localizing multiple bugs due to dealing only with the bug affects the accuracy of localization as stated by the authors. The graph represented here is undirected where the nodes of the graph represent the program statements and the edges represent the execution between them. Degree centrality is a graph centrality to measure the importance of a node in a network which will indicate that the part of the code will be more probable to contain an error. Another measure in which closeness centrality was used for each node to know the shortest path length between the node and other nodes. The result of this step will affect the process of multiple bug localization. The technique is evaluated on about 5 out of 7 programs from the Siemens dataset (Dset6, Dset6, Dset7, Dset8, Dset9, and Dset11). In the experiment on single fault localization, 99% of the faulty version can be found by exploring only 80% of the code. In the two bug's version, about 99% of bugs found after exploring 70% of the executable code. They evaluated their work based on the exam score evaluation and the incremental Developer Expense (IDE) methods.

Program slicing according to [12] [31] is a debugging technique that formulates a slice of code which are statements that affect a variable. Static slicing is a type of program slicing that generates slices depend on control dependencies in the code. Another type of program slicing is dynamic slicing which works on reducing the amount of space generated by static slicing. Dynamic slicing creates the slice depend on the variable values at run time to reduce the number of statements of the program in the debugging. However, execution slicing as stated by [32] applied data flow tests to formulate the slice or a group of slices (dice) by detecting the most probable statements from the tests to have the bug.

III. Motivational Example

This section presents the software artifacts of the software system explained within the motivational example. These artifacts will be later the input the application of different bug localization techniques in the following sections. The system description will be discussed in subsection “A”, a subset of system source code files will be presented in subsection “B”, and a subset of the software bug reports are shown in section “C”.

A. System Description

Consider a system for online shopping. The aim of the system is to be utilized for online shopping. The customer can browse some products, add them to his shopping cart then process the order. The order will be finalized, and the total amount will be calculated including taxes and the customer payment choice. The customer chooses a payment method and assigns it a profile as it is either cash, or by credit card, and the customer can update such payment method later. The administrator of the shop can add new computer products to the inventory with specific data. The shop has two main types of components: "DesktopLaptop” or “ComputerComponents". Also, the administrator can update taxes for any product, and products of the same type must be updated automatically.

Fig. 1 shows a partial class diagram the ‘Online Shopping System’ (OSS) including 9 classes. Customer class holds the customer’s information and operations does like adding a newproduct to the shopping cart (addProductToShopping () method) and assign a payment (setPayementMethod () method). ShoppingCart class holds information about products selected by the customer. Payment Method class is an interface for the type of payment, and it has two subclasses PaybyCredit and PaybyCache, with specific attributes for payment. ComputerProduct class is a parent class that consists of the basic information of any computer product of the system. DesktopLaptop and ComputerComponents are child classes of ComputerProduct class, each with specific properties. Inventory class manages the inventory through the addProduct () method for adding products with their quantity to the system. A relationship exists between Customer and ShoppingCart classes because each customer must have a shopping cart to add products to it. The relationship exists between Customer and PaymentMethod since each customer must decide his payment method for online shopping. ShoppingCart and Inventory classes are in an aggregation relationship with ComputerProduct class, as both classes consist of computer products. Such software system has a set of software artifacts that are presented within the following subsections.

B. Motivational Example Source Code Files

A subset of source code files for the online shopping system is presented in this section. The source code for the “Shopping Cart” class is presented in Fig. 2. The source code for the “Inventory” class is shown in Fig. 3.

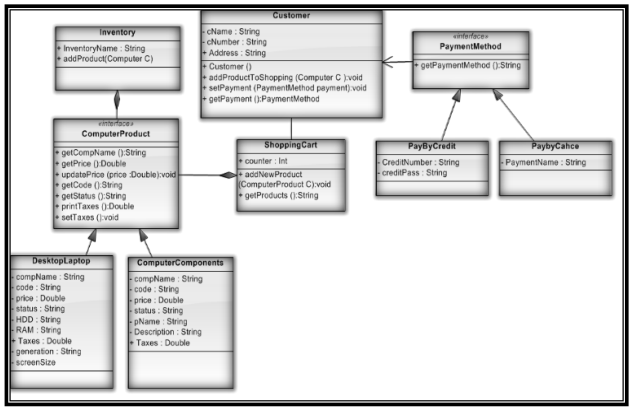


Fig. 1. A Partial Class Diagram of the Online Shopping System (OSS)

```
4       String inventoryName;
5       LinkedList<Computer>  comp ;
6       int counter = 0;
7       public Inventory(String inventoryName) {
8           this.inventoryName = inventoryName;
9           comp = new LinkedList<>();
10      }
11      public void addProduct(Computer c)
12      {
13          this.comp.add(c);
14      }
15  }
```

Fig. 2. "ShoppingCart .java" Source Code.

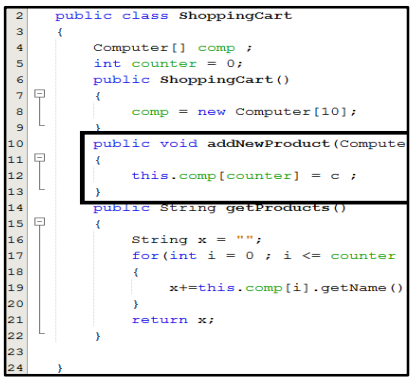


Fig. 3. "Inventory.java" Source Code File from File from the use Case Example.

C. Motivational Example Bug Reports

A bug report is a terminology that refers to documenting and describing software bugs that appeared while running a software [33]. As stated by [33] a bug report can be submitted by different stakeholders related to the software project such as the tester or the developer or user to the system. They posted their bug reports on a bug tracking system [33] which is used mainly for open-source projects to track different bug-report changes, assigned to solve the bug, or any other discussions.

Four bug reports, for the used motivational scenario, are presented in this section. Three bug reports have the status "resolved fixed" and one new bug report has the status "New".

TABLE I. BUG REPORT 1

Bug Report 1	
Bug ID	1102
Bug Summary	The payment method didn't change
Bug Status	Resolved Fixed
Product	Normal user
Reported	Online Marketing Application
Version	5/5/2020
Bug Description	I purchased pc and two other products then when I proceed to the order, they give me a note the payment will be on the cache given; However, I updated my payment method to pay by credit before.
Stack Trace	-
Fixed Files	Customer.java
Fixed Time	7/5/2020
Test Cases	-

Bug report 1, shown in Table I, shows a user who had previously changed his payment method from using cash to using the credit card. When making a new purchase afterwards, the system still displayed that his payment method will be using cash. Bug report 2, shown in Table II, has the status New" as it will be fixed by our example. The bug appears with the user when adding a new product to purchase to his cart, the program crashed and stopped. Bug report 3 shown in Table III presented a solved bug by adding a new product to the store. When the user of the system adds a new product to the system, a crash occurred. The bug was solved by the maintainers and the source code file "Inventory. Java". Bug report 4 shown in Table IV presented a solved bug with getting an invoice for a purchasing process. It was found that the tax percent is calculated incorrectly however it is calculated before. The solution to the bug is found in the source file "ComputerComponents. Java".

Starting from Section III to Section V, different bug localization techniques will be discussed and applied to the motivational scenario showing how those techniques work and their limitations.

TABLE II. BUG REPORT 2

Bug Report 2	
Bug ID	1104
Bug Summary	Adding a PC to purchase cause an error
Bug Status	New
Product	Online Marketing Application
Reported	5/12/2020
Version	1.2
Bug Description	When trying to add a PC to purchase and browse some other components and added them then adding another pc to the cart it is crashed.
Stack Trace	-
Fixed Files	-
Fixed Time	-
Test Cases	-

TABLE III. BUG REPORT 3

Bug Report 3	
Bug ID	1201
Bug Summary	Error with adding a new product to the store
Bug Status	Resolved Fixed
Product	Online Marketing Application
Reported	5/5/2020
Version	1.1
Bug Description	When I trying to add new product to the store, the program crashed given the following error
Stack Trace	Exception in thread "main" java.lang.OutOfMemoryError: GC overhead limit exceeded at java.util.LinkedList.linkLast(LinkedList.java:142) at java.util.LinkedList.add(LinkedList.java:338) at Inventory.addProduct(Inventory.java:28) at project. main(project.java:15)
Fixed Files	Inventory.java
Fixed Time	5/9/2020
Test Cases	1201

TABLE IV. BUG REPORT 4

Bug Report 4	
Bug ID	1325
Bug Summary	Error with getting an invoice
Bug Status	Resolved Fixed
Product	Online Marketing Application
Reported	5/5/2020
Version	1.1
Bug Description	When the transaction is going to be fired, it calculates the total invoice wrong with a problem in taxes percent
Stack Trace	-
Fixed Files	ComputerComponents.java
Fixed Time	7/5/2020
Test Cases	-

IV. INFORMATION RETRIEVAL TECHNIQUES

Information retrieval (IR) [34] is finding or extracting beneficial data that may be documents of unstructured nature like text that answers information needs. In the bug localization process, the source code files, bug reports either old or new, stack traces artifacts [35] will be the unstructured text of the system being analyzed. The unstructured data like the natural text in the bug reports, stack traces and source code file terms. This data needs to be retrieved and ranked using specific queries to retrieve the file contains bug [35]. The information retrieval passes through steps from preprocessing and preparing different text sources to similarity measures.

A. Case Study IR Experiment

In this subsection, three experiments will be applied. First, historical similar bug reports artifact will be utilized. Then source code artifact will be utilized in the second experiment. Finally, Similar Bug Reports Experiment applied.

1) Similar bug reports application: The first bug localization technique to apply, is an information retrieval technique that uses similarity scores across bugs. Klaus Changsun et al. [9] proposed a technique to localize bugs using an information retrieval technique. The assumption of their work depends on that if there is a new bug report similar in its attributes to one of the old bug reports then the fixed source code file by this old bug report will be the recommended source code file to be fixed with the new bug report. Such technique was applied to calculate similarity scores between the three resolved bug reports and the newly added bug report within the presented motivational example. The first step is to convert each bug report to a text vector as shown in Table V.

Then the Term Frequency Inverse Document Frequency (TF-IDF) measure [36] will be applied to the text of the bug reports. The calculated similarity measure between the new bug report (i.e., bug report 2) and each of the old bug reports resulted in the following scores presented in Table VI: bug report 1 is 0.2, bug report 3 is 0.16, and bug report 4 is 0.11. The experiment resulted in that bug report 1 is the most similar bug report to the new bug report. It means that the fixed file within bug report 1 (Customer .java) in the old bug report 1 is the file that contains the bug.

To evaluate the presented experiment, the new bug report needed to be fixed manually to know the files that contain the bug. The result of the manual investigation that the source code file "ShoppingCart.java". However, experiment 1 resulted in that the "Customer.java" is the file that contains the bug which means that the experiment 1 result is not true.

To understand why the applied bug localization technique failed to locate the source code file that contained the bug, a closer look is needed at the used bug reports. As per the bug's description in Section 3.2, bug report 1 was fixed by a change in Customer.java, whereas bug report 3 was fixed by a change in Inventory.java. The similarity score between bug report 1 and the new bug report was higher than the similarity score between bug report 3 and the new bug report. Hence, the applied bug localization technique suggested fixing the same file that was fixed previously by bug report 1. Hence, the applied technique could lead to a wrong location based on the text used within the newly opened bug. Such text is usually written by an end user, who has no knowledge of the inner details of the source code. Hence, relying on the text of the bug report solely is one main drawback of that bug localization technique. Another drawback is the complete reliance of the technique on the presence of historically fixed bug reports to recommend resolutions for the new bugs. Such assumption is not realistic when developing new applications that do not have a repository of previously fixed bug reports.

TABLE V. TEXT VECTORS OF THE BUG REPORTS OF OSS SYSTEM

Bug Report	*Bug Report Text Vector*
Bug Report 1	[Payment, Method, change, Normal, User, Application, purchased, two, products, proceed, order, they, give, note, payment, cache, given, updated, method, pay, credit]
Bug Report 2 (NEW)	[Adding, PC, purchase, cause, error, Online, Marketing, Application, trying, add, browse, some, components, added, adding, another, pc, cart, crashed
Bug Report 3	[Error, with, new, product, store, Online, Application, store, program, given, following, Exception, thread, main, javalangOutOfMemoryError, GC, overhead, limit, exceeded, at, javautilLinkedListlinkLastLinkedListjava142, javautilLinkedListaddLinkedListjava338, InventoryaddProductInventoryjava28, project, mainprojectjava15Inventoryjava]
Bug Report 4	[Error, new, product, store, Resolved, Fixed, transaction, going, fired, calculates, total, invoice, wrong, problem, taxes, percent]

TABLE VI. SIMILARITY SCORES BETWEEN THE NEW AND OLD REPORTS

Old Bug Report	**Similarity Score with the new bug report**
Old Bug Report 1	0.20
Old Bug Report 3	0.16
Old Bug Report 4	0.11

2) Source code experiment: In the second experiment, Similarity Scores between the new Bug report and the source code files are applied [9]. As experiment 1, similarity scores will be calculated. The difference here that text similarity will be applied between the new bug report and project source code files with the TF-IDF technique.

The number of source code files is nine files as listed in Table VII with their text vectors. In the same table, the similarity score between these sources code files and the new bug report is calculated. The similarity results must be sorted in descending order. But in this case, there are no common words between the new bug report and all source files.

TABLE VII. TEXT VECTORS OF THE BUG REPORTS WITH SOURCE FILES TEXT OF OSS SYSTEM

Source Code Files (.java)	Source File Text Vector	Similarity Score
ShoppingCart	[ShoppingCart, Computer, comp, counter, Computer, addNewGoodsComputer, compcounter, getGoods, xthiscompigetName, compigetPricen]	0
Inventory	[Inventory, inventoryName, Computer, comp, counter, Inventory, inventoryName, addProductComputer, comp, add]	0
Customer	[Customer, cName, cNumber, PaymentMethod, payment, ShoppingCart, PaybyCahce, addProductToShoppingComputer, shaddNewGoodsc, setPaymentPaymentMethod, thispayment, getPaymentMethod]	0
Computer	[Computer, getName, double, getPrice, updatePricedouble, price, getCode, getStatus, printTaxes, setTaxesdouble, taxes]	0
ComputerComponents	[ComputerComponents, implements, Computer, compName, code, price, status, pName, Description, taxes, ComputerComponentsString, thiscompName, code, price, thispName, Description, taxes, Override, getName, return, getPrice, getCode, getStatus, status, printTaxes, setTaxesdouble, updatePricedouble]	0
DesktopLaptop	[DesktopLaptop, implements, Computer, compName, code, price, status, HDD, RAM, generation, screenSize, taxes, DesktopLaptopString, compName, code, price, HDD, RAM, generation, screenSize, taxes, status, Offered, Override, getName, return, getPrice, getCode, getStatus, printTaxes, void, setTaxesdouble, updatePricedouble]	0
PaybyCahce	[public, class, PayByCredit, implements, PaymentMethod, Override, String, getMethod, return, enter, card, number, pass]	0
PayByCredit	[PaybyCahce, implements, PaymentMethod, Override, getMethod, Pay, cache]	0
PaymentMethod	[PaymentMethod, getMethod]	0

Discussion: As per the above similarity calculation, the text of the bug report does not match the naming conventions used within source files. Hence, relying on similarity scores analysis between the source code and bug reports would not result in locating bugs. The absence of such similarity is attributed to the constructing of those bug reports by a normal user who uses terms not related to the developer terms used within the source code files. So, the bug localization system in this state will not resulted in a true source code file. An example comparing the bug report called "NEW BUG REPORT 1" text to source file text as an example "ShoppingCart.java". The similarity scores between all the source code files and the new bug report equal to zero as no common words between them. After computing the same way with all source files, the new bug report got zero similarity score with all of them.

3) Stack traces: Stack traces or execution traces represent the method calls during the execution of the application. When an error occurs during the execution and the program stops working or works in an unexpected way, the current state of the stack trace represents the method calls till the stopping point.

The presence of stack traces, as a part of the bug report, will enhance the accuracy of finding the source code file that contains the error [37]. From an information retrieval perspective, having a stack trace as a part of the bug report will result in higher similarity score between the bug reports and the source code files. Furthermore, stack traces result in faster manual debugging by the developers [38]. For example, Fig. 4 shows represents a bug from eclipse [39] how the file names involved in the error and the corresponding line number are shown appear or the line that contains an error are shown in Fig. 4 that line 13 contains the error in the source file inventory.java. Schroter et al. apply [40] a study on 3940 bug reports. 2,321 bugs reports observed that they are fixed contains stack traces with 60 %. Also, the mean lifetime of the bugs include stack traces is 2.73 Days compared with the remaining bug report does not contain stack traces with mean 4.13 days. So, in our case, bug report 2 with status new will not be solved using stack traces as the bug report does not contain it.

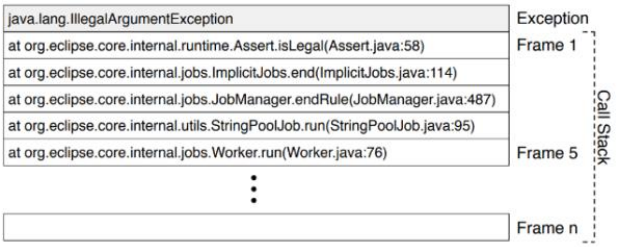


Fig. 4. Sample Stack Trace Extracted from ECLIPSE Bug.

V. MACHINE LEARNING TECHNIQUES

Machine learning is a branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment [41]. Different bug localization techniques utilized different machine learning techniques to localize bugs automatically.

A. Case Study Experiment using Machine Learning

The most important step in applying a machine learning algorithm is the preparation of the features. For the motivational scenario, the features will be similarity scores between each old bug report and all the source code files. Each record or row of features represents this similarity scores with the source code files as presented in the following Table IX. Three bug reports will be utilized from the motivational example. Two of them will be used for training and one for testing. Their calculated similarity scores with the source code files to be the features. Also, the result of these features which the source file contains the error with each bug report appeared in the right cell of the row. As these bug reports are solved before and the files contains solved error already known. After preparing the training set with these two bug reports, it will be fed to the machine learning algorithm. Then the testing phase started by preparing the features for a bug report that we know its result before. The new bug report is prepared with its specific features. Then the machine learning algorithm will decide its decision which appeared in the last row in Table VIII. As shown in Table VIII, we have only two training examples with only two results. After running a machine learning algorithm, with feeding the training examples to the machine learning. Then feeding the testing example as to tell us the source code file containing the error. The result will be "ComputerComponents .java".

The file that contains the bug for new bug report 4 will be “ComputerComponents.java”. That means the result of the experiment is not true. Different reasons lead to such a wrong location of the bug. First, machine learning needs a huge training set to learn, otherwise it will not work properly [42]. Second, if the project has no old, solved bugs, machine learning will not be an applicable technique. In such a case, the only alternative would be to take training data from a different software project, like projects similar in nature to make use of their old bug reports. Some challenges will face this work: the language of the project may be different, the type of project as it may be desktop, web application or other.

TABLE VIII. FEATURES PREPARATION FOR MACHINE LEARNING ALGORITHM BUG REPORT

Training example	**Feature 1 (ShoppingCart)**	**Feature 2 (Inventory)**	**Feature N**	**Result (Source File)**
Training example 1 (*Bug report 1*)	0.11	0.1	…..	*Customer*
Training example 2 (*Bug report 3)*	0.3	0.1	0.3	*Inventory*
Testing example (Bug Report 4)	0.1	0.5	0.1	*?*

VI. PROGRAM SPECTRUM TECHNIQUES

Program spectra refer to the program entities that are covered during the execution of the program [29], [43]. Also, the spectrum based get some information executed from the programs as the test cases. There are several types of spectra [29] used in the spectra based fault localization as (program statements, variables, execution trace, execution path, path profile, execution profile, Number of failed test cases cover a statement and not, number of successful test cases that cover a statement and not, the total number of test cases that cover a statement, the total number of test cases that do not cover a statement, the total number of successful test cases, total number of failed test cases and the test case number. Such technique demands the presence of test cases, or the presence of correct program execution traces, to be applied. Furthermore, the technique demands having a large set of test cases to cover the lines of code that most probably contains an error.

A. Case Study Experiment using Program Spectrum

The main inputs to this experiment will be the test cases. Some operations must be applied to test cases to know the lines of code that will most probably have the error. Two test cases related to the motivational example are shown in Tables IX and X. When a new bug is reported, the output of the test cases (i.e., the spectrum) can be utilized with some equations to find the bug [44]. The source code file that the test cases will run on is presented in Fig. 2.

TABLE IX. TEST CASE 1 FOR OSS SYSTEM

Test Case 1	
Test Case Number	1
Test Case Inputs	Create Object of Inventory
Expected Output	Object Created without error
Actual Output	Created Successfully
Test Case Result	Success

TABLE X. TEST CASE 2 FOR OSS SYSTEM

Test Case 2	
Test Case Number	2
Test Case Inputs	Add new Product to Inventory
Expected Output	Added Successfully
Actual Output	Added Successfully
Test Case Result	Success

From the test cases and different inputs, test case number (1) will execute the following lines 6 to 13. And test case number (2) will cover 14 and 16. Then the number of records will be counted for each line to find the bug. The technique applied for spectrum depends on the statistical equation for every line of code covered by test cases. In Table XI, different program spectrum listed in section 2.2.3 as number of successful test cases that execute lines of code are presented with the occurrence of each spectrum with each line of code that are in the vertical rows. The hit of a spectra with a line of code represented by 1 and 0 for nit hitting. The criteria for each line like the number of successful test cases covered, the number of failed test cases, overall test cases in each line. We will consider that we have only two test cases for our bug localization task. Test case one presented in the above figure, it will hit the class source code from line 6 to 13 and the result of this test cases is a success. Test case two will hit the lines of

code from 14 to 16. Consider Line 6 as shown in the table for illustration, we must calculate different spectrum for each line from the resulted test case as follows: Number of success test cases (NCS)covered line 6 will be test case number (1) only so the total will be one. The number of failed test cases (N_{CF}) covered in line 6 will be equal to zero as our test cases here are only two and both are successful. The number of test cases covers line 6 is equal to one as we list it before. Several test cases not covered in line 6 are equal to one which is test case number (2). The number of failed test cases not covered (NUF) line 6 is equal to zero as we have only two successful test cases. The last spectrum is the Number of Success Test Cases Not Covered which is one as test case 2 is a successful test case not covered line 6. The total number of failed test cases (NF) =0. The above steps will be calculated to all lines of the code as shown in the table. Then an equation is applied to calculate different spectra in one number for all lines of code then we have to sort these scores in descending order. The highest score will represent the line that contains the error.

The equation performed here that utilized different spectrum used from [44] presented in the (1):

$$Suspiciousness(Ochiai) = \frac{N_{CF}}{\sqrt{N_{F}*(N_{CF}+N_{CS})}} \quad (1)$$

The score to line 6 will be equal to NCF = 0, NF=0, NCS=1 by substituting, the result will equal to zero. The above process will be repeated for every line of code then sorted but here all scores equal to zero.

The main limitation of program spectrum is that the many test cases need to be analyzed, to find the bug then many test cases to be tested to find a true solution which affects the time and performance of bug localization process [45]. Unfortunately, the results are equal, also we need to compute many test cases that affect the time to find the source code file contains bug [45].

TABLE XI. DIFFERENT PROGRAM SPECTRA FOR A SOURCE CODE FILE FOR OSS SYSTEM

Different Program Spectra	**Code Line**									
	6	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Success Test Cases Covered	1	1	1	1	1	1	1	1	1	1
Failed Test Cases Covered	0	0	0	0	0	0	0	0	0	0
Test Cases Covered	1	1	1	1	1	1	1	1	1	1
Test Cases not Covered	1	1	1	1	1	1	1	1	1	1
Failed Test Cases not Covered	0	0	0	0	0	0	0	0	0	0
Success Test Cases not Covered	1	1	1	1	1	1	1	1	1	1
Total Score	0	0	0	0	0	0	0	0	0	0

VII. DISCUSSION AND CONCLUSION

This paper presents a review to explore different software bug localization techniques. The exploration done through presenting different past works. Additionally, a motivational example is applied to show how these techniques are working presenting their limitations; also, the software artifacts that are utilized and which are not utilized.

Two main findings are presented: (1) Some software artifacts are not properly utilized in the process of software bug localization. (2) The current software bug localization techniques suffer from some limitations. We elaborate on those findings as follows.

Finding 1: Many bug localization systems use information from both bug reports and source files. Previous research [10], [46], [47], [48], [49], [50], [51], [15] utilized the natural text of the bug reports with terms of the source files. Method names and the abstract syntax trees are used from source code files [10]. However, the changes that applied to each source code file among different from version control systems used by [49], [15]. Also, application interface descriptions text has been utilized by [49] , [50]. Test cases are also used where successful test cases and failing test cases are used to find the most probable error [22] [11] [26].

However, several artifacts are not utilized in the bug localization process. They are software requirements, use cases, classes' relationships within the source code, software architecture, and different comments between developers or written discussion between them of old bug reports that may affect the process of localization mentioned. If we have bug report text data as stated above. Text data can be linked to requirements text, which can be then, shorten the search with source code files to specific files of a definite module.

Finding 2: Different software bug localization techniques are applied in the process of bug localization (Information Retrieval, Machine Learning, and Program spectrum). These techniques suffer from some limitations and this appeared from applying the motivational example.

Information retrieval techniques suffer from the problem of the dependence on natural unstructured text. Those techniques depend on matching the new bug report text to any of the old bug reports, and hence recommending the fixed file of the old bug report. But such technique may not take us to the true old bug report depending on how the bug report is written, which varies greatly between developers and end users of the system. This issue appears as well if we attempt to measure the text similarities between the new bug report and the source code files. Also, the lack of old bug reports for the same project may prohibit applying the technique altogether.

Machine Learning techniques will not work properly in two situations. In the first situation, that we train the bug localization system on some software projects, and the new bug appears in a different project. Hence, the bug localization system will not give the exact source code that contains the error. The past works [46] [47] [48] [49] [11] are applied on one of the datasets and tested on the same dataset. The second situation when we do not have enough training examples to train the machine learning algorithm.

Test case-based techniques as program spectrum and program slicing are depending on test cases to localize bugs. The main limitation comes with performance and time to test the whole system to find the bug. Also, the huge number of test cases is to be examined to find the bug.

Accordingly, we anticipate that by utilizing additional software artifacts, and additional information from previously utilized software artifacts, we can improve the accuracy of bug localization, and extend its applicability even to projects that do not have historical information about the source code of the fixed bugs.

REFERENCES

[1] S. M. H. Dehaghani and N. Hajrahimi, "Which factors affect software projects maintenance cost more?," Acta Informatica Medica, vol. 21, no. 1, p. 63, 2013.

[2] L. Erlikh, "Leveraging Legacy System Dollars for E-Business," IT Professional , vol. 2, no. 3, pp. 17-23, 2000.

[3] H. B., T. B. and M. K., "Software Maintenance Implications on Cost and Schedule," in 2008 IEEE Aerospace Conference, 2008.

[4] B. P. Lientz and E. B. Swanson, Software maintenance management, Addison-Wesley Longman Publishing Co., 1980.

[5] I. Sommerville, Software Engineering, Addison-wesley, 2007.

[6] A. Kumar and B. S. Gill, "Maintenance vs. reengineering software systems," Global Journal of Computer Science and Technology, vol. 11, no. 23, 2012.

[7] D. Cubrani´c, "Automatic bug triage using text categorization," in the International Conference on Software Engineering & Knowledge Engineering, Alberta, 2004.

[8] W. W. Eric, G. Ruizhi, L. Yihao, R. Abreu and W. Franz, "A survey on software fault localization," IEEE Transactions on Software Engineering, pp. 707-740, 2016.

[9] K. Youm, J. Ahn and E. Lee, "Improved bug localization based on code change histories and bug reports," Information and Software Technology, pp. 177-192, 2017.

[10] S. LU, W. MEILIN and Y. YUXING, "Deep Learning With Customized Abstract Syntax Tree for Bug Localization," IEEE Access 7 , vol. 7, pp. 116309-116320, 2019.

[11] K. Jeongho, K. Jindae and L. Eunseok, "VFL: Variable-based fault localization," Information and Software Technology, p. 179–191, 2019.

[12] T. Frank, " A survey of program slicing techniques," Journal of Programming Languages, vol. 3, pp. 121-189, 1995.

[13] J. Zhou, H. Zhang and D. Lo., "Where should the bugs be fixed? more accurate information retrieval-based bug localization based on bug reports," in Software Engineering (ICSE), 2012 34th International Conference on, IEEE, 2012.

[14] Z. Wen, L. Ziqiang, W. Qing and L. Juan, "FineLocator: A novel approach to method-level fine-grained bug localization by query expansion," Information and Software Technology, pp. 1-15, 2019.

[15] W. Yaojing, Y. Yuan, T. Hanghang, X. Huo, L. Ming, X. Feng and L. Jian, "Bug Localization via Supervised Topic Modeling," in 2018 IEEE International Conference on Data Mining (ICDM)., 2018.

[16] C. Mills, P. Jevgenija, P. Esteban, B. Gabriele and H. Sonia, "Are Bug Reports Enough for Text Retrieval-based Bug Localization," in 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME), 2018.

[17] Y. Zhou, Y. Tong, R. Gu and H. Gall, "Combining text mining and data mining for bug report classification," Journal of Software: Evolution and Process, vol. 228, no. 3, pp. 150-176, 2016.

[18] M. D'Ambros, M. Lanza and R. Robbes, "An extensive comparison of bug prediction approaches," in Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on. IEEE, 2010.

[19] A. Murgia, G. Concas and M. Marchesi, "A machine learning approach for text categorization of fixing-issue commits on CVS," in the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (p. 6). ACM., 2010.

[20] A. N. Lam, A. T. Nguyen, H. A. Nguyen and T. N. Nguyen, "Combining deep learning with information retrieval to localize buggy files for bug reports," in 2015 30th IEEE/ACM International Conference on Automated Software Engineering, 2015.

[21] D. Kim, Y. Tao, S. Kim and A. Zeller, "Where should we fix this bug? a two-phase recommendation model," IEEE transactions on software Engineering, vol. 39, no. 11, pp. 1597-1610, 2013.

[22] W. ERIC and Q. YU, "BP neural network-based effective fault localization," International Journal of Software Engineering and Knowledge Engineering, vol. 19, no. 4, pp. 573-593, 2009.

[23] J. A. Jones and M. J. Harrold, "Empirical evaluation of the tarantula automatic fault-localization," in the 20th IEEE/ACM international Conference on Automated software engineering, 2005.

[24] Z. Ziye, L. Yun, W. Yu, T. Hanghang and W. Yaojing, "A deep multimodal model for bug localization," Data Mining and Knowledge Discovery, vol. 35, no. 4, pp. 1369-1392, 2021.

[25] B. Qi, S. Hailong, Y. Wei, Z. Hongyu and M. Xiangxin, "DreamLoc: A Deep Relevance Matching-Based Framework for bug Localization," IEEE Transactions on Reliability , 2021.

[26] R. Henrique, d. A. Roberto, C. Marcos, S. Higor and K. Fabio, "Jaguar: a spectrum-based fault localization tool for real-world software," in 2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST), 2018.

[27] A. Rui, Z. Peter, G. Rob and G. Arjan, "A practical evaluation of spectrum-based fault localization," The Journal of Systems and Software, p. 1780–1792, 2009.

[28] V. B´ela, "NFL: Neighbor-Based Fault Localization Technique," in IEEE 1st International Workshop on Intelligent Bug Fixing (IBF), 2019.

[29] S. Higor, C. Marcos and K. Fabio, "Spectrum-based Software Fault Localization: A Survey of Techniques, Advances, and Challenges," arXiv preprint arXiv:1607.04347, 2016.

[30] Z. Abubakar, P. ,. Sai and Y. C. Chun, "Simultaneous localization of software faults based on complex network theory," IEEE Access, pp. 23990-24002, 2018.

[31] M. Weiser, "Programmers use slices when debugging," Communications of the ACM, vol. 25, no. 7, pp. 446-452, 1982.

[32] W. Eric and D. Vidroha, "A Survey of Software Fault Localization," Department of Computer Science, University of Texas at Dallas, Tech. Rep. UTDCS-45 9, Texas, 2009.

[33] J. Zhang, X. Wang, D. Hao, B. Xie, L. Zhang and H. Mei, "A survey on bug-report analysis," Science China Information Sciences, vol. 58, no. 2, pp. 1-24, 2015.

[34] U. Cambridge, Introduction to information retrieval, 2009.

[35] S. Wang and D. Lo, "Amalgam+: Composing rich information sources for accurate bug localization," Journal of Software: Evolution and Process, pp. 921-942, 2016.

[36] A. Aizawa, "An information-theoretic perspective of tf--idf measures," Information Processing & Management, vol. 39, no. 1, pp. 45-65, 2003.

[37] W. Shaowei and L. David, "Amalgam+: Composing rich information sources for accurate bug localization," Journal of Software: Evolution and Process , vol. 28, p. 921–942, 2016.

[38] N. Bettenburg, R. Premraj, T. Zimmermann and S. Kim, "Extracting structural information from bug reports," in The 2008 international working conference on Mining software repositories, 2008.

[39] S. Adrian, B. Nicolas and P. Rahul, "Do Stack Traces Help Developers Fix Bugs?," in 2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010), 2010.

[40] S. Adrian, B. Nicolas and P. Rahul, "Do stack traces help developers fix bugs?," in 2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010), 2010.

[41] M. Sayed, R. K. Salem and A. E. Khder, "A Survey of Arabic Text Classification Approaches," International Journal of Computer Applications in Technology, vol. 95, no. 3, pp. 236-251, 2019.

[42] K. J. Haider and Z. K. Rafiqul, "Methods to Avoid over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study)," Computer Science, Communication & Instrumentation Devices, pp. 163-172, 2015.

[43] T. Reps, T. Ball, M. Das and J. Larus, Software Engineering—Esec/Fse'97, Springer, 1997.

[44] A. Rui, Z. Peter and J. Arjan, "An evaluation of similarity coefficients for software fault localization," in 12th Pacific Rim International Symposium on Dependable Computing (PRDC'06), 2006.

[45] V. László and B. Árpád, "Test suite reduction for fault detection and localization: A combined approach.," in 014 Software Evolution Week-IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering , 2014.

[46] H. Xuan, T. Ferdian, L. Ming, L. David and S. Shu-Ting, "Deep Transfer Bug Localization," IEEE Transactions on Software Engineering, pp. 1-12, 2019.

[47] X. Yan, K. Jacky, M. Qing and B. Kwabena, "Bug Localization with Semantic and Structural Features using Convolutional Neural Network and Cascade Forest," in of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018. ACM, 2018.

[48] H. Xuan, L. Ming and Z. Zhi-Hua, "Learning Unified Features from Natural and Programming Languages for Locating Buggy Source Code," in IJCAI, 2016.

[49] Y. Xin, B. Razvan and L. Chang, "Learning to Rank Relevant Files for Bug Reports using Domain Knowledge," in In Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2014.

[50] N,. An, T. N. Anh, A. ,. Hoan and N. N. Tien, "Combining Deep Learning with Information Retrieval to Localize Buggy Files for Bug Reports," in 30th IEEE/ACM International Conference on Automated Software Engineering, 2015.

[51] S. Jeongju and Y. Shin, "FLUCCS: Using Code and Change Metrics to Improve Fault Localization," in the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis, 2018.

Failure Region Estimation of Linear Voltage Regulator using Model-based Virtual Sensing and Non-invasive Stability Measurement

Syukri Zamri, Mohd Hairi Mohd Zaman, Muhammad Fauzi Mohd Raihan
Asraf Mohamed Moubark, M Marzuki Mustafa
Department of Electrical, Electronic and Systems Engineering
Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia

***Abstract*—Voltage regulator (VR) stability plays an essential role in ensuring maximum power delivery and long-lasting electronic lifespan. Capacitor with a specific equivalent series resistance (ESR) range is typically connected at the VR output terminal to compensate for instability of the VR due to sudden changes in load current. The stability of VR can be measured by analyzing output voltage during load transient tests. However, the optimum ESR range obtained from the ESR tunnel graph in its datasheet can only be characterized by testing a set of data points consisting of ESR and load currents. Characterization process is performed manually by changing the value of ESR and load current for each operating point. However, the inefficient process of estimating the critical value of ESR must be improved given that it requires a large amount of time and expertise. Furthermore, the stability analysis is currently conducted on the basis of the number of oscillation counts of VR output voltage signal. Therefore, a model-based virtual sensing approach that mainly focuses on black-box modeling through system identification method and training neural network on the basis of estimated transfer function coefficients is introduced in this study. The proposed approach is used to estimate the internal model of the VR and reduce the number of data points that need to be acquired. In addition, the VR stability is analyzed using noninvasive stability measurement method, which can measure phase margin from the frequency response of the VR circuit in closed-loop conditions. Results showed that the proposed method reduces the time it takes to produce an ESR tunnel graph by 84% with reasonable accuracy (MSE of 5×10^{-6}, RMSE of 2.24×10^{-3}, MAE of 1×10^{-3}, and R^2 of 0.99). Therefore, efficiency and effectiveness of ESR characterization and stability analysis of the VR circuit is improved.**

Keywords*—Voltage regulator; output capacitor; equivalent series resistance; failure region; system identification; neural network; noninvasive stability measurement

I. Introduction

Increasing demand for electronic products, such as system-on-chips and personal electronics, commonly requires the use of a voltage regulator (VR) for stable and regulated output voltage supply. VR has been widely used in the electronic field due to the development of new technologies and increasing demand for high-performance electronic devices and compact solutions [1],[2]. VRs in electronic devices are embedded in an integrated circuit (IC), but fault probability of the VR can be influenced by a few parameters, such as temperature, input voltage supply, and aging factors [3],[5],[6]. These factors may further deteriorate internal parameters and thus reduce the performance of electronic devices or completely eliminate their functionality [21]. Therefore, industrial electronic manufacturers must perform stability analysis of the VR.

The existing analysis for VR stability through ESR and load is performed manually [4],[7]. This method is solely conducted by testing a vast number of data, observing load transient, and varying the load current for a specific ESR value. Thus, an accurate ESR tunnel graph can be obtained to show stable and unstable regions for operating conditions of the VR. This situation occurs because an internal model for the VR is lacking and product variations may cause parameters inside the VR to vary. Therefore, analyzing VR stability without prior knowledge of the VR internal model is challenging [9],[10]. Additionally, variation of load currents may also cause VR instability and inefficiency [9],[21]. Hence, an efficient and accurate failure region estimation method is necessary under the condition that the actual model is known.

A. VR Mechanism

The two different types of VRs are linear (LVR) and switching VRs. LVRs are low cost and can regulate a small drop-out voltage with less noise compared with switching VRs [5],[6],[11]. Hence, LVR can minimize the amount of power loss in the internal VR and is highly efficient. VRs aim to regulate the input voltage supply and produce low-noise, constant, and stable output DC voltage [4],[5], thereby indicating the absence of multiple oscillations or ripples. Moreover, VRs can limit over- and undershoot values during sudden changes in the load current.

As shown in Fig. 1, a typical VR circuitry contains an output capacitor connected at the output terminal that acts as an energy storage element. Moreover, the output capacitor compensates for the disturbance during load transient [7],[8]. However, impurity element inside the capacitor called equivalent series resistance (ESR) is a main factor that contributes to the stability of the VR. Although a pure capacitor should ideally contain only the capacitance value without ohmic resistance, the case is different in the real world.

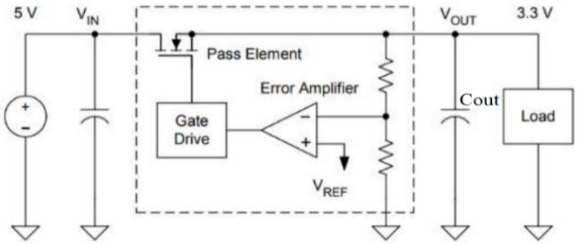


Fig. 1. Basic LVR Circuit.

B. ESR Tunnel Graph

The optimum value of ESR is crucial in determining the stability of the VR. An excessively high or low value of ESR may cause a significant output voltage to undershoot, produce unwanted oscillations, and cause instability. The ESR compensates for the disturbance by adding zero to its transfer function to cancel out the non-dominant pole and thus achieve a dominant-pole compensation.

Manufacturers provide a datasheet for each fabricated VR to depict the stable range of ESR values in the VR circuit through a unique chart called the ESR tunnel graph. Fig. 2 shows an example of an ESR tunnel graph using the TPS76301, a commercial VR from Texas Instrument. The plot presents a range of ESR values against a range of load currents. A specific ESR and output current range is chosen for plotting the ESR tunnel graph. The figure shows that the value of ESR from 0.3 Ω to 10 Ω indicates a stable region while other values denote non-stable regions. The critical value of ESR is located at the failure region boundary.

C. VR Stability Analysis

LVR stability can be analyzed using two types of responses: (a) load transient response in the time domain or (b) frequency response in the frequency domain. Stability analysis based on the load transient test is usually conducted because of its simplicity and the method can be performed under closed-loop condition despite its low accuracy [6]. Although the frequency response of the LVR can be ideally obtained when the system is under open-loop condition, this scenario is difficult to achieve in the actual case. The LVR system is typically packaged under closed-loop condition; therefore, yielding its transient response is simple [17]. Furthermore, the frequency response can yield a more accurate stability measurement than the transient response because it indicates the phase margin of the system [13],[18],[20]. However, determining the frequency response is challenging because it can only be obtained while the system is under open-loop condition and this scenario breaks the loop in the actual LVR. Thus, a method called noninvasive stability measurement (NSM) is proposed to obtain the frequency response of the VR system under closed-loop condition.

Studies on VR stability based on the NSM method are limited. Recent studies typically analyze the electronic system through transient response [5], [6], [16]. However, investigations based on the frequency response are few. Existing studies mainly focus on fault diagnosis [3], [12], [19], scalability, and dynamic performance [12] but those on achieving short-time stability analysis of VR are limited.

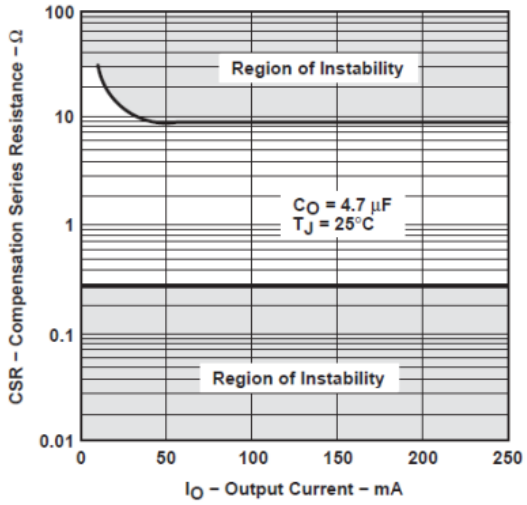


Fig. 2. Example of the ESR Tunnel Graph (VR Model TPS76301 from Texas Instruments).

NSM can be utilized to improve the existing stability characterization method in the VR failure region estimation process and determine the stability condition of a particular operating point. In addition, a virtual sensing approach that enhances the black-box modeling method can be used to illustrate the model of the internal VR circuitry and estimate the model transfer function coefficient used to determine critical ESR values located in the failure region boundary. Therefore, a model-based virtual sensing approach (MBA) that mainly focuses on the black-box modeling through system identification (SI) and training the neural network (NN) on the basis of the estimated transfer function coefficient is introduced in this study, as explained in Section II. MBA is used to estimate the internal model of the VR and reduce the number of data points required to estimate the critical value of ESR for generating the region of failure of the VR stability through the ESR tunnel graph. Furthermore, outcomes from this MBA approach are described and discussed in Section III.

II. Methods

Four phases of this study is presented in Fig. 3. The first phase is the manual characterization, which analyzes the load transient test of the VR. The outcome of this phase is also used as the benchmark for the proposed method in this work. The second phase implements the NSM method to obtain the phase margin of the VR circuit in each operating point in the ESR tunnel graph. The third phase applies the MBA by first estimating the VR system model using the system identification method and then training the neural network structure. The final phase validates the method performance using various performance metrics.

A. VR Manual Characterization as Benchmark

The commercial LVR used in this study is the LT1963A from Analog Devices because of the comprehensive information provided in its datasheet [14] and its availability in the LTSpice software for simulation purposes. The LVR circuitry developed for the manual ESR characterization with a step signal is illustrated in Fig. 4. This step signal is used to

disturb the load current in the load transient test. The 10 µF capacitor used in this work is based on the datasheet provided by manufacturers. A resistor is connected in series with the output capacitor and labeled ESR given that ESR is absent in the purely capacitive capacitor used in the simulation. As mentioned in the early section, VR characterization is performed manually in manufacturing practice. Hence, the ESR of the output capacitor manually varies and the undershoot, overshoot, and oscillations during the load transient test are observed for stability analysis.

The datasheet also indicated that the range of the input voltage should be between 2.5 and 20 V when obtaining the output voltage range of 1.21–20 V [14]. Therefore, two resistors (R1 and R2) must be chosen appropriately to obtain an output voltage of 5 V. Values of R1 and R2 can be calculated as follows:

$$V_{out} = V_{adj}\left(1 + \frac{R_2}{R_1}\right) + \left(I_{adj}\right)(R_2), \quad (1)$$

where Vadj is 1.21 V and Iadj is 3 µA. Therefore, values obtained for R1 and R2 are 12 and 3.9 k, respectively.

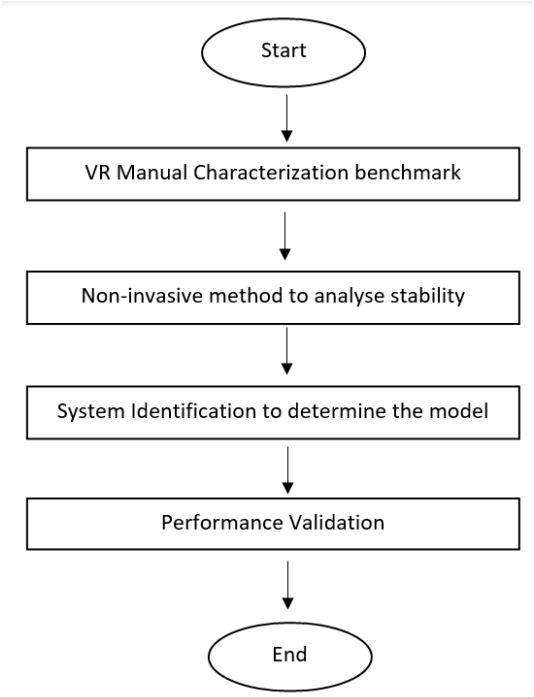


Fig. 3. Flowchart of the Proposed Method.

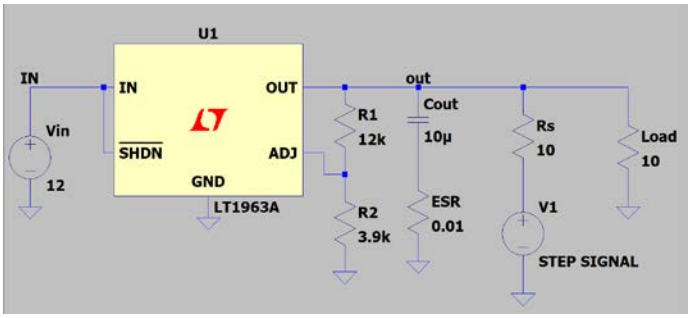


Fig. 4. LT1963A VR Circuit in LTSpice for Load Transient Test.

The ESR value is manually changed from 0.01 Ω to 0.3 Ω, with an increment of 0.01 Ω. Meanwhile, the load current is in the range of 0.01–0.05 A, with an increment of 0.01 A. Combining each ESR value and each load current produces one operating point. The load transient test is then conducted. The circuit is energized to obtain the transient response after the input voltage is initialized and the ESR and load current values are configured and set to a specific value. The start time is recorded immediately after the process begins until every operating point in the ESR tunnel graph is tested.

The stability of each operating point is determined manually on the basis of the output voltage observed in the load transient test response during the manual characterization. As stated in the LT1963A datasheet, the ESR value must be between 20 mΩ and 3 Ω for an output voltage of 1.2 V with a 10 µF output capacitor to ensure VR stability [14]. The output voltage oscillation must be examined for each operating data point in this case. Otherwise, the VR system is considered unstable with excessive ringing, that is, more than three oscillations exist. This stability condition check is also performed manually and requires high expertise. Manual characterization is conducted on all data points. Finally, an ESR tunnel graph is illustrated to depict stable and unstable ranges of ESR for a specific load current.

B. Noninvasive Stability Measurement

The proposed stability measurement method is based on the NSM method, which analyzes the VR stability under closed-loop condition to obtain the phase margin of the VR system. Fig. 5 shows the LVR circuitry setup to obtain the phase margin from the frequency response of the LVR system using a small-injection AC signal with an injection transformer at the output terminal of the LVR.

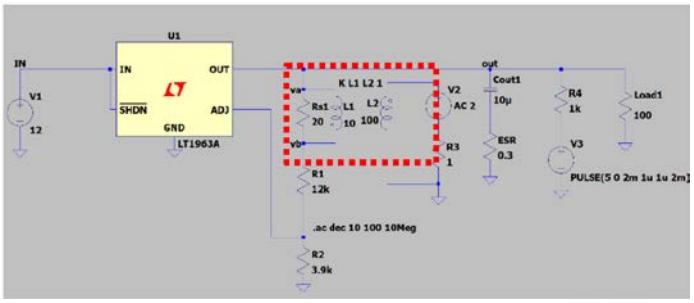


Fig. 5. LVR Circuitry for Noninvasive Stability Measurement.

The NSM method is proposed to obtain the phase margin through frequency response without breaking the loop condition of the system. Thus, the result obtained from this approach is accurate and efficient because breaking the control loop of the VR system is unnecessary. A small-injection resistor Rinj with a value of 20 Ω, which is relatively small compared with that of R1 and R2, is connected in series at the upper terminal of R1 to perform the NSM method and ensure that two different injection points (v_a and v_b) can be established. A ground (GND) reference for these points is absent; therefore, an injection transformer is connected parallel to Rinj on the primary side *L1*. Meanwhile, the secondary side *L2* is connected to a sine wave signal generator V2. Both points v_a and v_b are then connected to an oscilloscope to ensure

that both sine wave signals entering the system at point v_b and exiting the system at point v_a can be observed. Amplitude gains of both sine wave signals are expected to differ; thus, frequency tuning is required until both sine waves display the same gain, which is equal to 1 or also known as unity gain. The frequency when both sine wave channels demonstrate the same gain is known as crossover frequency at 0 db or unity gain frequency. Another parameter that must be considered is the phase shift between the two signal waves at points v_a and v_b. The phase shift value between the two signal waves represents the phase margin of the system. Therefore, the ESR tunnel graph can be produced by observing the system's frequency response for each operating data point tested using the NSM method and an accurate stability condition is expected.

C. VR System Modeling Through SI

The next step is to apply the black-box modeling approach through SI to estimate the VR circuit model. The SI method is used to estimate the internal model of the VR circuit given that input and output data are available [15]. The circuit used for SI data acquisition is similar to the one displayed in Fig. 4. However, the voltage source V1 in Fig. 4 generates a pseudorandom binary signal (PRBS) instead of a step signal in SI. Steps taken in the SI approach are the preprocessing of data, estimation and validation of data, model structure determination, and choosing the desired coefficient of the model. Hence, the model of the VR system can be evaluated with the optimal fitness model.

Input, output, and sampling time must be determined prior to data preprocessing. In this case, input data are the small-signal output voltage V_{out} and output data are the small-signal output current I_{out}. These data are obtained from the circuit simulation using LTSpice by exporting all data into the MATLAB software. Further processing is then conducted in MATLAB. The removal of the mean value of raw data after importing raw data is also known as detrend. Half of detrended data is used for estimation data while the other half is utilized for validation data.

The SI model structure selection must be determined for estimating the VR model. Several types of model structures, such as autoregressive exogenous input (ARX), output–error (OE), autoregressive moving average exogenous input (ARMAX), and Box–Jenkins (BJ) model structures, can be used to estimate the model of a dynamic system [7], [8]. This work utilized the OE model structure due to its simpler model transfer function parameters compared with those of other model structures. Fig. 6 shows the output–error model structure.

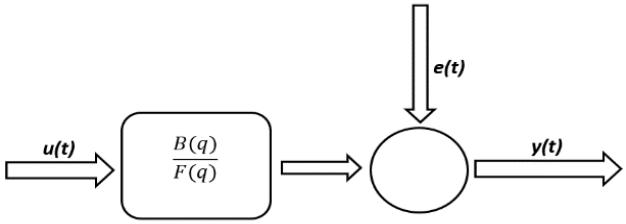


Fig. 6. OE Model Structure.

The OE model structure can be expressed as follows:

$$y(t) = \frac{B(q)}{F(q)}u(t - n_k) + e(t) \quad (2)$$

where *y(t)* is the model output, *u(t)* is the model input, n_k is the number of delays, e_k is the white noise error, and *k* is the number of samples. In addition, the polynomial B(q) represents the numerator in relation to the input *u(t)* and F(q) represents the denominator in relation to the output *y(t)*. Polynomials B(q) and F(q) can be expressed with the backward shift operator term q^{-1} as follows:

$$B(q) = \sum_{k=1}^{\infty} b_k q^{-k} = b_1 + b_2 q^{-1} + \cdots + b_{n_b} q^{-n_b+1}, \quad (3)$$

$$F(q) = \sum_{k=0}^{n_f} f_k q^{-k} = 1 + f_1 q^{-1} + \cdots + f_{n_f} q^{-n_f}, \quad (4)$$

where n_b is the order of polynomial B(q) and n_f is the order of polynomial F(q). The following process is used for model estimation using the linear regression for an iterative method with unknown parameters θ:

$$y(k, \theta) = \phi(k^T)\theta = \frac{B(q)}{F(q)}u(t) = \xi(k, \theta), \quad (5)$$

where ϕ(k) is expressed as

$$\phi(k) = [u(k-1), u(k-2), \ldots, u(k-n_b),$$
$$-\xi(k-1, \theta), -\xi(k-2, \theta), \ldots, -\xi\big(k - n_f, \theta\big)], \quad (6)$$

where θ is expressed as

$$\theta = [b_1, b_2, \cdots, b_{n_b}, f_1, f_2, \cdots, f_{n_f}]^{\boldsymbol{T}} \quad (7)$$

Therefore, the percentage error of actual output data and the estimated output model can be reduced by obtaining the model transfer function coefficient or parameter vector θ. We then apply validation data to the estimated transfer function. Hence, the model fitness can be obtained and the transfer function with the maximum percentage of the fitness model is selected. Validation of the selected SI model transfer function is continued with a step signal that changes from 0 V to 5 V after a slight delay. The transient response from SI is recorded and then compared with the one in the LVR circuitry simulation during the load transient test to validate the SI-estimated model.

D. Neural Network Training

The following process shows the training of the NN structure to reduce the number of operating points by testing the few sets of operating data points. Therefore, VR characterization time can be significantly reduced with the decrease of testing of operating data points. Fig. 7 shows an example of an NN structure consisting of input, hidden, and output layers with a number of neurons.

Input data are fed into the NN structure via channels, which are typically assigned with numerical values and known as the weight, to the hidden layer in this stage. Input layers are then multiplied to their own corresponding weights, and the hidden layer performs its mathematical computation. The output layer predicts the output, which is the estimated model transfer function coefficients previously obtained from the SI. Finally, the ESR and load currents are fed into the input layer while

transfer function coefficients of the SI model are fed into the output layer. These input selections are chosen due to their correlation with the coefficient of the output transfer function for each operating data point in the ESR tunnel graph.

The trained NN structure was then used to estimate the model transfer function for the remaining untested operating data. The step response was obtained through MATLAB simulation for each operating data point after estimating all transfer function coefficients using the trained NN. Step responses from both manual characterization and MBA (SI-NN) are then compared and validated for their similarity. Finally, an ESR tunnel graph from the MBA-based characterization is produced and then compared with the ESR tunnel graph from the manual characterization in terms of critical ESR values.

E. Performance Validation

The last stage evaluates the obtained ESR critical values from both manual and MBVS characterization processes and determines the efficiency of the MBVA characterization method compared with the manual process. Mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient (R^2), and efficiency calculation can be expressed as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left[y(i) - y_p(i)\right]^2, \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left[y(i) - y_p(i)\right]^2}, \quad (9)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|y(i) - y_p(i)\right|, \quad (10)$$

$$R^2 = \frac{\sum_{i=1}^{n}[\left(y(i)-\bar{y}(i)\right)\left(y_p(i)-\bar{y}_p(i)\right)]}{\sqrt{\sum_{i=1}^{n}(y(i)-\bar{y}(i))^2 \sum_{i=1}^{n}(y_p(i)-\bar{y}_p(i))^2}}, \quad (11)$$

$$Efficiency = \left[1 - \frac{t_{SI-NN}}{t_{manual}}\right] \times 100\%, \quad (12)$$

where the term y is the actual critical ESR value, y_p is the critical ESR value obtained from the proposed method, n is the number of observations, and i is the number of load current instants. The output of the proposed method is the SI-NN characterization and validated if MSE, RMSE, and MAE values are close to zero. The efficiency value determines how the time is taken for the proposed method to be conducted compared to the manual characterization method.

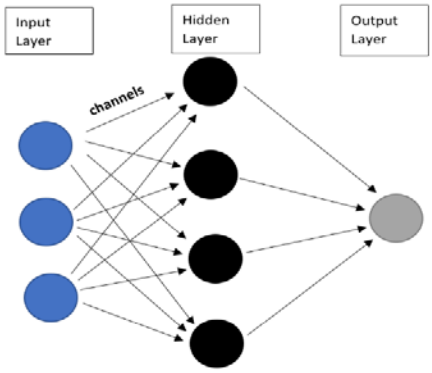


Fig. 7. Neural Network Structure.

III. Results and Discussion

A. Manual VR Characterization Results

This stage phase aims to produce a benchmark for the ESR tunnel graph that depicts stable and unstable regions of operating data points. As mentioned earlier, the ESR value must be higher than 20 mΩ for an output voltage of 1.2 V to ensure VR stability. The load transient of the LVR circuit is simulated and then the transient response is observed to analyze its corresponding number of oscillations at this operating condition. Fig. 8 shows the transient response obtained from the load transient for an LVR circuit with an output current of 500 mA, ESR value of 20 mΩ, and output voltage of 1.2 V. As shown in Fig. 8, the number of oscillations obtained is three cycles with an undershoot of 31.34 mV.

A voltage drop of 31.34 mV is observed from the first wave of the load transient. Stability analysis is carried out using the noninvasive method under closed-loop conditions after the load transient is obtained from the circuit simulation for each operating data point to provide increasingly accurate and efficient stability measurement through the system's frequency response. Fig. 9 depicts the Bode plot to obtain the phase margin of the system through the frequency domain. Component parameters of the circuit for this noninvasive method are the same as those used to obtain the load transient test circuit.

The phase margin obtained from the noninvasive method was 17.64° at a crossover frequency of 206.28 kHz. This phase margin value indicates the border region of the system stability. Hence, an ESR tunnel graph depicted in Fig. 10 is the product of all operating data points tested using the noninvasive method.

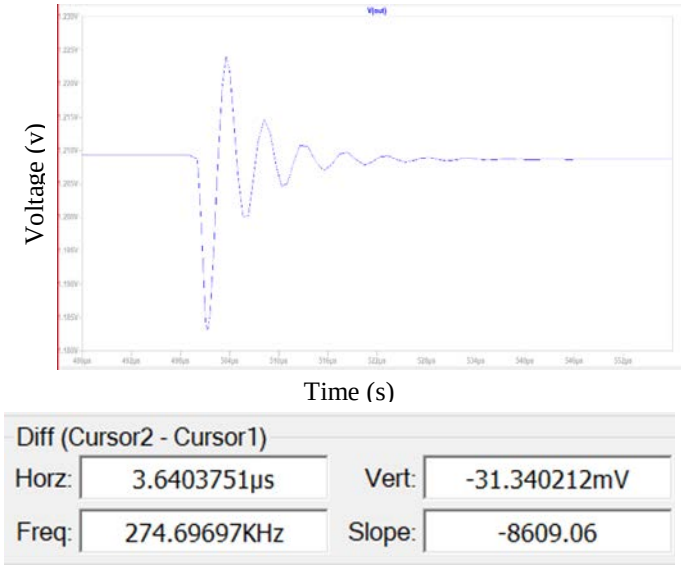


Fig. 8. Transient Response for an Output Current of 500 mA with an ESR of 20 mΩ.

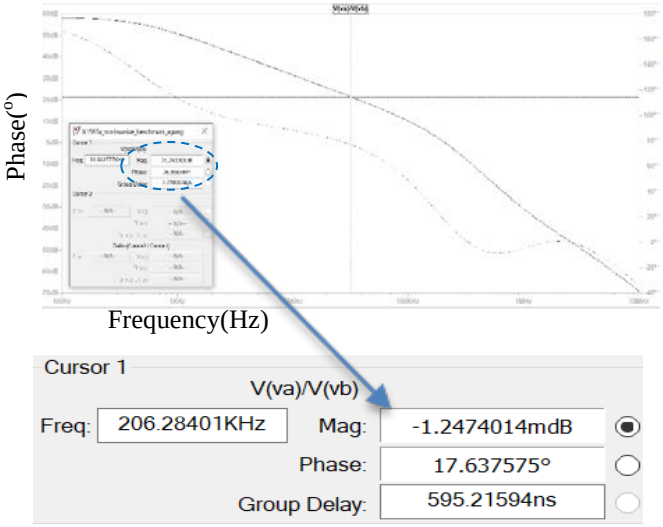


Fig. 9. Phase Margin Results using the Noninvasive Method.

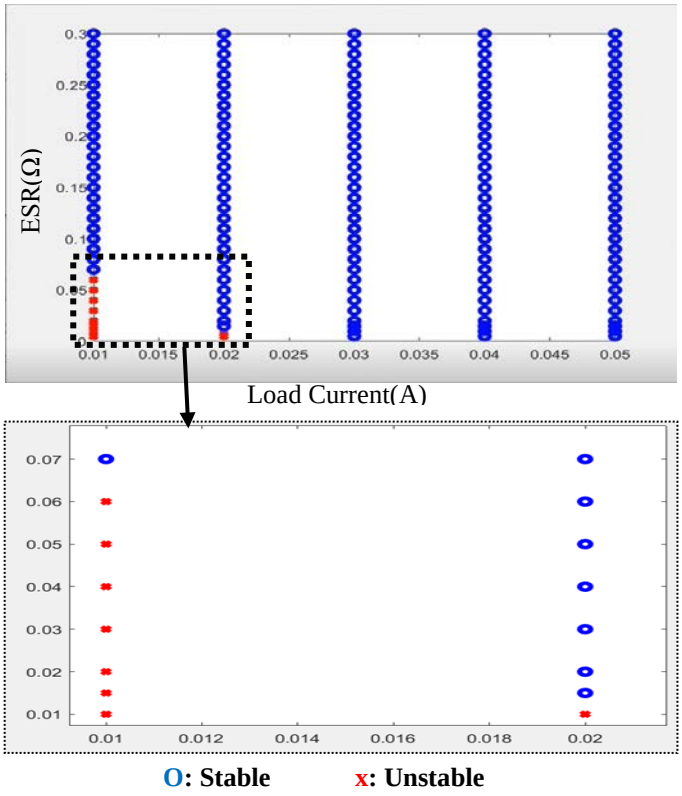


Fig. 10. ESR Tunnel Graph through Manual Characterization.

B. SI-NN Characterization Results

The SI-NN characterization method is approached after the manual benchmark has been obtained. First, the SI method uses the black-box modeling concept to determine the transfer function coefficient that represents the system model. Second, the model selection with the maximum percentage of model fitness is selected to represent the system model. Fig. 11 shows the output model fitness by tuning parameters of poles and zeroes of the OE models. The percentage fitness of different model parameters of OE is presented in Table I.

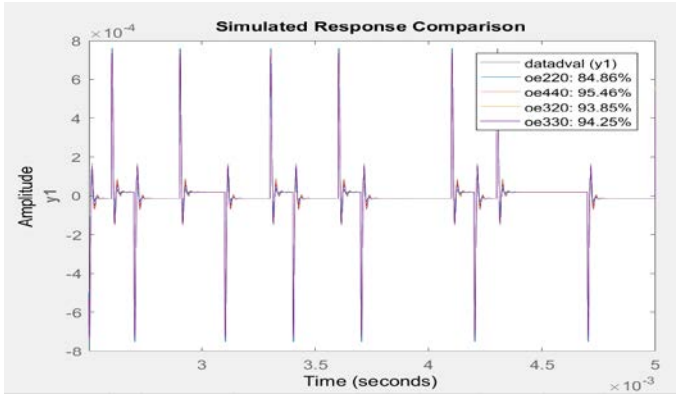


Fig. 11. Output Models with Fitness Percentage.

TABLE I. FITNESS PERCENTAGE OF DIFFERENT MODELS

Model	Fitness Percentage (%)
OE220	84.86
OE320	93.85
OE330	94.25
OE440	95.46

The model fitness percentage with the minimum number of parameters is chosen due to its simplicity. Thus, the OE320 model is selected. The output yields a transfer function coefficient for *B(q)* and *F(q)* parameters on the basis of this model (Fig. 12). Finally, the transfer function coefficient is tabulated and used for NN training and the dataset reduction phase.

```
oe320 =
Discrete-time OE model:  y(t) = [B(z)/F(z)]u(t) + e(t)
  B(z) = -0.09031 + 0.1189 z^-1 - 0.02923 z^-2

  F(z) = 1 - 1.601 z^-1 + 0.7026 z^-2
```

Fig. 12. Transfer Function Coefficient for OE320.

All individual coefficients are applied to the output of the NN structure after the transfer function coefficient obtained from all operating data points is characterized and tabulated through SI to validate the output model. All individual transfer function coefficients are combined and fed into the output layer of the NN structure. Therefore, the value of the NN output layer is 5. Finally, as mentioned earlier, the ESR and output current are fed to the input layer of the NN. Thus, the input layer is 2, and the hidden layer varies from 10 to 50 with an increment of 10. Fig. 13 shows the network architecture of an NN structure trained for the selected estimated model.

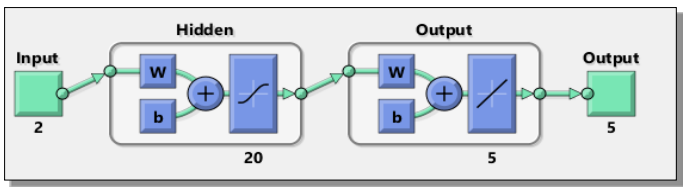


Fig. 13. NN Structure.

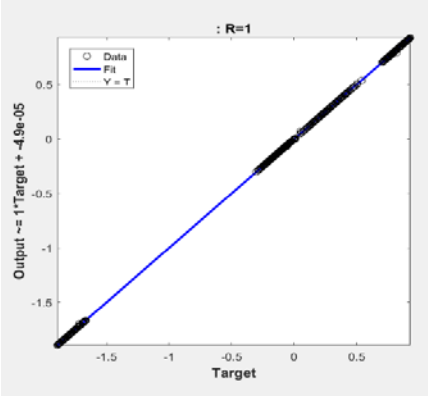


Fig. 14. Regression Plot.

The dataset used for training into the NN varies by reducing the percentage of the total dataset from the SI beginning with 93.7%, 87.5%, 81.25%, 75%, and 68.75%. For each of the reduced percentage dataset reductions, a different number of hidden layers, as mentioned earlier, is assigned to train the neural network structure starting. Bayesian regularization (BR) is then used as the training algorithm. Fig. 14 illustrates the regression plot of the output data obtained. The value of R^2 indicates the correlation between measured and target outputs. A value approaching 1 indicates a close and precise relationship.

C. ESR Tunnel Graph using SI-NN

The ESR tunnel graph benchmark from manual characterization is then compared with the ESR tunnel graph obtained using the SI-NN characterization, with the phase margin as the targeted output. Fig. 15 shows the ESR tunnel graph obtained from the SI-NN approach.

D. Performance Metrics

Performance metric parameters in Tables II and III were observed for a different number of dataset reductions and a fixed hidden layer size of 20 and 10, respectively, to validate the results of the SI-NN characterization method further.

The calculated metrics showed that at 20 number of trained data, for hidden layer size of 20, yields the most negligible MSE value of $5x10^{-6}$ that showed a high critical ESR value prediction.

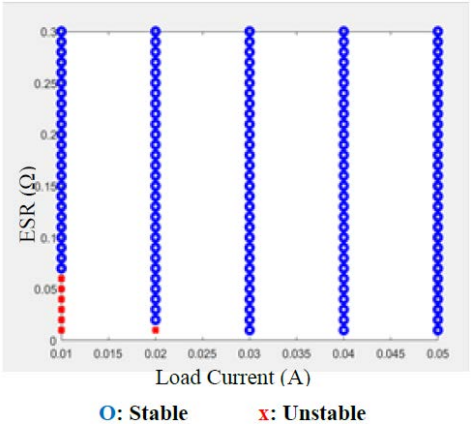


Fig. 15. ESR Tunnel Graph using the SI-NN Approach.

TABLE II. PERFORMANCE METRICS FOR DIFFERENT NUMBERS OF REDUCED TRAINED DATA AT A HIDDEN LAYER SIZE OF 20

No. of trained data	Performance Metrics			
	MSE	RMSE	MAE	R^2
10	5×10^{-5}	7.07×10^{-3}	2×10^{-3}	0.987
20	5×10^{-6}	2.24×10^{-3}	1×10^{-3}	0.999
30	4.5×10^{-5}	6.71×10^{-3}	3×10^{-3}	0.999
40	5×10^{-6}	2.24×10^{-3}	1×10^{-3}	0.999
50	5×10^{-6}	2.24×10^{-3}	1×10^{-3}	0.999

TABLE III. PERFORMANCE METRICS FOR DIFFERENT NUMBERS OF REDUCED TRAINED DATA AT A HIDDEN LAYER SIZE OF 10

No. of trained data	Performance Metrics			
	MSE	RMSE	MAE	R^2
10	6.85×10^{-4}	2.62×10^{-2}	1.5×10^{-2}	0.99
20	2.5×10^{-4}	5×10^{-3}	3×10^{-3}	0.91
30	5×10^{-6}	2.24×10^{-3}	1×10^{-3}	0.99
40	5×10^{-6}	2.24×10^{-3}	1×10^{-3}	0.99
50	5×10^{-6}	2.24×10^{-3}	1×10^{-3}	0.99

IV. CONCLUSION

The proposed method can generally reduce the amount of time taken to characterize the failure region of the voltage regulator, and estimate critical ESR values that accurately distinguish stable and unstable regions of the voltage regulator system. The proposed method can estimate the internal model of VR through the SI method. Furthermore, the VR output voltage stability can be determined via a noninvasive stability measurement approach without breaking the internal control loop inside the VR circuit.

ACKNOWLEDGMENT

The authors acknowledge the financial support received from the Ministry of Higher Education Malaysia through research grant no. FRGS/1/2019/TK04/UKM/03/1.

REFERENCES

[1] M. H. Jahanbakhshi & M. Etezadinejad. 2019. Modeling and current balancing of interleaved buck converter using single current sensor. *27th Iranian Conference on Electrical Engineering (ICEE2019)*, pp. 662-667.

[2] O. Garcia, P. Zumel, A. de Castro, P. Alou & J. Cobos. A. 2008. Current self-balance mechanism in multiphase buck converter. *2008 IEEE Power Electronics Specialists Conference, 2008*, pp. 624-628.

[3] I. Kovacs, M. Topa, M. Ene, A. Buzo & G. Pelz. 2020. A metamodel-based adaptive sampling approach for efficient failure region characterization of integrated circuits. *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*, pp. 1-5.

[4] M. H. M. Zaman, M. M. Mustafa, M. A. Hannan & A. Hussain. 2018. Neural network based prediction of stable equivalent series resistance in voltage regulator characterization. *Bulletin of Electrical Engineering and Informatics*, vol.7, no.1, pp. 134-142.

[5] N. Sedaghati, H. Martinez-Garcia & J. Cosp-Vilella. 2016. On modeling of linear assisted DC-DC voltage regulators. *2016 Conference on Design of Circuits and Integrated Systems (DCIS)*, pp. 1-4.

[6] X. Ming, H. Liang, Z. W. Zhang, Y. L. Xin, Y. Qin & Z. Wang. A High Efficiency and Fast-Transient Low-Dropout Regulator With Adaptive Pole Tracking Frequency Compensation Technique. *IEEE Transactions on Power Electronics*, vol.35, no.11, pp. 12401-12415, 2020.

[7] M. H. M. Zaman, M. M. Mustafa & A. Hussain. 2018. Estimation of voltage regulator stable region using radial basis function neural network. *Journal of Telecommunication, Electronic and Computer Engineering*, vol.10, no.2-8, pp.63-66.

[8] M. H. M. Zaman, M. M. Mustafa & A. Hussain. 2017. Critical equivalent series resistance estimation for voltage regulator stability usinghybrid system identification and neural network. *International Journal on Advanced Science, Engineering and Information Technology*, vol.7, no.4, pp.1381-1388.

[9] C. Wang, C. Huang, T. Lee & U. F. Chio. 2006. A linear LDO regulator with modified NMCF frequency compensation independent of off-chip capacitor and ESR. *APCCAS 2006-2006 IEEE Asia Pacific Conference on Circuits and Systems.* pp. 880-883.

[10] M.Day. 2002. Understanding low drop out (LDO) regulators. *Texas Instruments, Dallas.* pp. 1-6.

[11] N. Tang, Y. Tang, Z. Zhou, B. Nguyen, W. Hong, P. Zhang, J. H. Kim & D. Heo. 2018. Analog-assisted digital capacitorless low-dropout regulator supporting wide load range. *IEEE Transactions on Industrial Electronics* 2019, vol. 66, no. 3, pp. 1799-1808.

[12] K. Laadjal & M. Sahraoui. 2020. On-Line fault diagnosis of DC-Link electrolytic capacitors in boost converters using the STFT technique. *IEEE Transactions on Power Electronics* 2021, vol. 36, no. 6, pp. 6303-6312.

[13] M. Ho, J. Guo, K. H. Mak, W. L. Goh, S. Bu, Y. Zheng, X. Tang & K. N. Leung. 2016. A CMOS low-dropout regulator with dominant pole-substitution. *IEEE Transactions on Power Electronics*, vol. 31, no. 9, pp. 6362-6371.

[14] LT1963A series: 1.5A, low noise, fast transient response LDO regulators data sheet. California, United States of America.

[15] T. Souvignet, T. Coulot, Y. David, S. Trochut, T. Di Gilio & B. Allard. 2013. Black box small-signal model of PMOS LDO voltage regulator. *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*, pp. 495-500.

[16] Y. Li, H. Fan, Q. Feng, Q. Hu, L. Hu, H. Chen & H. Heidari. 2020. A fast transient response and high PSR low drop-out voltage regulator. *27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1-4.

[17] B. Dennis & B. S. Kariyappa. 2018. Fault isolation in analog circuits using multi-support vector neural network. *3rd International Conference on Communication and Electronics Systems (ICCES)*, pp. 655-660.

[18] B. Dennis & B. S. Kariyappa. 2018. Support vector neural network and principal component analysis for fault diagnosis of analog circuits. *2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1152-1157.

[19] C. Yang, X. Zhang, A. He & L. Qiu. 2017. Fault diagnosis of analog circuit based on complex model. *32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 949-952.

[20] Y. K. Cho & B. H. Park. 2015. Loop stability compensation technique for continuous-time common-mode feedback circuits. *Internationol SoC Design Conference (ISOCC)*, pp. 241-242.

[21] M. Dobler, M. Harrant, M. Rafaila, G. Pelz, W. Rosenstiel & M. Bogdan. 2015. Bordersearch: An adaptive identification of failure regions. *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2015, pp. 1036-1041.

Mobile Mathematics Learning Application Selection using Fuzzy TOPSIS

Seren Başaran[0000-0001-9983-1442, 1]
Near East University, Department of Computer Information Systems
Lefkoşa 98010 via: Mersin 10 Turkey, Cyprus

Firass El Homsi[2]
Arab International University
Department of Management Information Systems
Damascus, Syria

***Abstract*—Impressive evolution of technology increased the usage frequency of smart mobile phones, and hence abundance in the quantity of available mobile applications has emerged as the vital problem of inventing practical and efficient ways for selecting suitable mobile applications for the desired use. Today, there are almost three million apps only at the Google Play store. Therefore, the need for an automated, effective, and less time-consuming approach towards suitable mobile application selection to choose the best alternative has gained more significance than ever. Despite the sudden growth in mobile learning applications, there exists a dearth of research in the effective way of selecting a suitable mobile application in that respect particularly in relation to mobile apps for Mathematics. Moreover, using multi-criteria decision-making methods (MCDM) is only recently applied in rare studies for that purpose. This paper focused on ISO/IEC 25010 software quality standards in selecting mobile Mathematics learning applications. Six highly rated applications were evaluated by two experts. This paper aims to apply the fuzzy Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to retrieve the best alternative among present applications. The results showed an objective and flexible assessment for ranking to eliminate ambiguity in decision-making. Results also identified significant features thus rendering a useful and valuable tool for decision-makers. The study assists users, teachers/instructors, students in their decision-making processes regarding finding the most suitable application for Mathematics.**

***Keywords*—Fuzzy TOPSIS; ISO/IEC 25010 standards; mathematics; mobile applications; multi-criteria decision making**

I. Introduction

The number of mobile applications developed each year is growing exponentially. This increase is highly noticeable in the area of digital learning objects [1]. These developments are justifiable by ideas given by [2] which contend that new, better, and effective learning platforms are needed to facilitate learning. It is also clearly seen that developments in mobile learning applications are concurrent with educational developments. In such regard, one can contend that ideas in [3] expressed gratitude to the development of mobile mathematics learning applications. This stems from their contribution towards learning algebra, statistics, geometry, mathematical analysis, and other calculations. The authors of the study in [4] also established that mobile learning applications have made it easy to harness meta-cognitive abilities and represent thoughts in a better way. Researchers in [5] posts that mobile learning applications are essential in dealing with matters that involve a lot of problem-solving and critical thinking. These contributions made by using mobile learning applications are numerous and some are continuing to be discovered with time. This is one of the major reasons why it is important to conduct studies related to the use of mobile learning applications especially in the field of mathematics.

Meanwhile, there exist different ways which can be used to determine the quality of mobile learning applications as well as their contributions towards improving user experience. These standards include ISO-25010, ISO-9126, and FURPS (Functionality, Usability, Reliability, Performance and Supportability). They primarily focus on the software quality aspects of the mobile learning applications [6], [7]. But most of the existing studies that evaluate the quality of mobile learning applications highly focus on technical aspects.

High-quality and improved user experience is not guaranteed by the availability of numerous alternative mobile learning applications for mathematics. The outcome of numerous researches shows that some of the existing mobile learning applications have not contributed much to learning [8], [9]. This is highly true with regards to observations which exhibit that quite a number of mobile learning applications for mathematics have high ratings which do not match their contributions towards improving learning [10].

The study conducted [6] strongly argues that some users prefer not to use mobile Mathematics learning applications (MMLAs). The primary reason is that they are not easy to use. Another reason was given by [3] which contend that several MMLAs involve a lot of manual selection. This is critical because it increases the time users spend before starting to have final access to the application. Matters are worsened by the fact that there is a lot of dissatisfaction surrounding the use of MMLAs. This is attributed to ideas which contend that MMLAs are not much different from traditional mathematics learning methods [6], [9]. That is, they are of low quality and do not contribute much towards improving user satisfaction.

As a result, it can thus be seen that there is a huge need to develop and select high-quality, user-friendly, and user-enhancing experience MMLAs. Another challenge posed is that this topic is a new and emerging one in the study of MMLAs. Hence, more work is needed to study how mathematics learning quality and user experience can be improved notably by using multi-criteria decision-making methods. These techniques can guarantee a high level of

success in quality evaluation [11]. For this purpose, this study thus seeks to use a fuzzy TOPSIS method to select high-quality and user-enhancing experiences of mobile learning applications for mathematics.

The main objective of this study is to use the fuzzy TOPSIS approach to select high-quality and user-enhancing experiences of mobile learning applications for mathematics. This study also seeks to identify problems that are undermining the use of MMLAs and offer solutions to deal with such challenges.

The study aids in enhancing the use and effectiveness of MMLAs. The outcomes of the study are believed to improve learning across all platforms by addressing significant challenges affecting the use of MMLAs. This study also provides standards in relation to ISO practices upon which the quality and usability of MMLAs can be determined. Moreover, it plays an important role to the study of mobile applications through the use of Fuzzy TOPSIS.

That is, it contributes towards improving existing empirical frameworks on Fuzzy TOPSIS. This technique is otherwise known as the Fuzzy TOPSIS approach and came into existence as a result of efforts to provide a framework for choosing the ideal alternative [12]. The decision is made based on the most and closest distance between the negative ideal solution and the positive ideal solution [13].

The rest of the paper is organized as follows: The literature review is discussed in Section II. The decision-making approach for evaluation is described in Section III with the introduction of sample mobile learning applications. The fourth section is dedicated to the findings and the subsequent discussion. The conclusion is discussed in Section V.

II. Related Work

There are circumstances that require a lot of decisions to be made simultaneously. These decisions are most times conflicting and require an objective criterion to help make the best and relevant decision. This leads to Multi-Criteria Decision Making (MCDM). MCDM primarily includes a combination of expert views and the use of historical data to make decisions [14]. It thus quantifies subjective judgments and implies that the best MLAM must be part of MCDM. MCDM is composed of a number of approaches which include: VIKOR, TOPSIS, ELECTRE, PROMOTHEE, and Analytic Hierarchy Process (AHP). The key to making the right decision is to examine the weight of the choices which vary in relation to their relative values. Hence, the challenges of subjectivity and imprecision are bound to be encountered in any MCDM activity. MCDM applies in a lot of circumstances and areas. In Mathematics, the MCDM can be related to quality, usability, costs, convenience, reliability, accuracy and dependability, accessibility, etc. For students to highly benefit from the use of MMLAs, these MCDM aspects are essential and must be prioritized.

The growth and advancement in technological accessibility and digital globalization over the past two decades have greatly impacted learning. This has opened more opportunities for learning by removing physical limitations. Hence, learning is becoming increasingly mobile [15]. In general terms, a mobile application is any software application that permits the user to undertake certain tasks through the use of a mobile (handheld or wireless) device such as tablets, smartphones. This promotes accessibility at any point in time in respect to the usual assigned position or location.

Using MLAs on handheld and wireless devices help promote collaboration and individualization of the learning process as hinted by [16]. This, therefore, means that students can learn at their own convenience and pace. The importance of this is heightened by the fact that learning is not hampered when students do not have physical access to classroom materials or desktop computers.

According to [17], mobile learning (m-learning) and electronic learning (e-learning) are greatly influenced by the existence of mobile learning devices (MLDs). Hence the absence of MLDs can hamper both m-learning and e-learning. This is further seen by insights provided by [18]. These insights established that the lack of MLDs hindered the growth, development, and use of mobile.

WELCOME (Wireless E-Learning and Communication Environment) was used by [19] as part of m-learning strategies to examine students' experience and performance. The results showed that m-learning is a desirable and essential feature for contemporary education. It further established that m-learning enhances the experience and effectiveness of students. This was further improved with the integration of WAP (Wireless Access Protocols).

Other researchers such as in [20] have focused on the integration of m-learning with SMS technology in universities. The study involved the use of whiteboards. In this study, students asked questions, took part in classroom discussions, and provided feedback. This feature greatly showed the need for proper categorization of students into the entire learning process by time, receiver, sender, etc. Such can also be extended to the examination of MMLAs. The most interesting development was by [17] and it involved the development of a tutoring system that allows users to access it using handheld and wireless devices. Such a system captured student performance, records and included an assessment platform. In addition, this has been a solid platform upon which MMLAs and other learning applications have been developed.

From all these insights, deductions can be made that mobile learning is an innovative approach to learning. This inference stems from the numerous benefits that users or learners are bound to get from using them. Mobile learning applications can thus be said to enhance convenience, accessibility, speed, interaction, collaboration activities in learning.

However, this relies on quite a several factors such as the availability of internet access and mobile devices. Also, researchers in [20] established that lack of quality can hinder the use of MLAs. Researchers in [19] noted the need to enhance user satisfaction as another key aspect to enhancing the use of MLAs. These issues are the driving motivation for this study to identify the high quality MMLAs using the Fuzzy TOPSIS approach.

The integration of MLAs in mathematics is a great innovative move that works towards improving learners' knowledge and understanding of mathematical aspects. Researcher in [21] posits that the use of MMLAs enhances learners' chances of being successful or performing better in mathematics. As such, the whole process of learning mathematics can be casual and unconstrained as users can use any MLDs such as cell phones and tablets.

It was highlighted that MMLAs tend to deal with arithmetic problems faced by learners [3]. This is because MMLAs are designed to suit any individual irrespective of his or her mathematical abilities and most of the modules provided start from elementary aspects or basics of any mathematics subject. Hence, MMLAs can be considered to deal with deeper mathematical issues such as numerical programming, critical thinking, geometrical constriction, charts representation, etc.

Authors in [22] used Maths4Mobile to look at the use and importance of arranged and social learning angles in learning mathematics. Their results provided support of the additional benefits obtained from using MMLAs over traditional learning methods. The cited reasons pointed towards increased coordination and engagement amongst the students.

A study that examined the situational learning environment involved the use of Nokia mobile phones to learn Mathematics [23]. Findings showed that the use of mobile phones greatly encouraged students to participate in learning mathematics. In addition, more students were observed to have greatly improved in their academic performance with regards to mathematics. Recommendations were made that the use of mobile phones encourage unaided learning and hence using MMLAs can play the same role too.

The use of MMLAs attracted and continues to attract the attention of major and reputable bodies which are in support of their use. For instance, the U.S. National Council of Teachers of Mathematics in 2008 encouraged educational institutions to allow students' access to MLAs. Such developments were said to foster speed, creativity, and innovation in learning[1].

The authors of the study in [24] gave different arguments concerning the use of MMLAs citing that they can also obstruct the learning process. This is considerably true as students can shift focus towards non-educational activities on mobile applications [25]. Despite the occurrence of these problems, it is still being advocated that MMLAs play an important role in mathematics [26]. Hence, we can expect such a notion to play an important role in learning mathematics as innovative developments continue to take place in the foreseeable future.

It is worthy to note that user satisfaction and quality enhancement are also important aspects to look at when examining both the importance and drawbacks of using MLAs. For instance, the use of MLAs does not guarantee user satisfaction. Such can be seen in reviews that are given by users who sometimes complain of using the MLAs. Hence, the number of users using the MLAs is often a good indicator of determining if such MLA is good or bad and if it has problems or not. Ratings are also another strategy that can be used to examine the existence of drawbacks. That is, higher ratings such as 4.5 and 5 or possibly more offer an indication that the MLA has little or no problems affecting it.

Though MMLAs learning applications have a lot of benefits that users can obtain from using them, they are still prone to suffer or pose numerous drawbacks. For instance, researchers in [6] established that most MLAs always fail to live up to expectations.

The reason is that they fail to serve the intended purpose. That is, not all MMLAs offer the desired mathematics learning materials and some materials are relatively few and inaccessible.

Meanwhile, applications are themselves part of the full composition of what is termed software and hence any problem that is surrounded by the use of software can affect the use of MLAs. For instance, software crush problems can make MLAs inaccessible and this can happen most when users are in great need of the application. Most of them require constant updates and may not work with certain mobile devices. For instance, certain MLAs are restricted to IOS while other work only on Android and Windows operating systems.

From all these drawbacks, the development of high quality and user enhancing MLAs has to consider all these challenges. As a result, an assumption can be made that mobile applications that have higher ratings such as 4.5 and 5 or possibly more and a high number of users, offer an indication that the MLA has little or no problems affecting it. However, multiple MLAs might have high ratings despite their flaws. To minimize this ambiguity, multi-criteria decision-making methods were offered.

Researchers examined the use of fuzzy TOPSIS and FAHP in addressing user satisfaction and quality issues involved in using MMLAs [6]. The study focused on 5 MMLAs with higher user ratings of 5 available on Google Play Store. The findings revealed that the best and less time-consuming MMLAs can be selected by using Fuzzy TOPSIS and FAHP.

Fuzzy TOPSIS approach is better when used to rank the decisions while the FAHP works better in assigning weights [27]. This entails that the Fuzzy TOPSIS approach works more efficiently in ranking the best MMLAs.

The study also used the Fuzzy TOPSIS and FAHP to analyze the food industry's product life cycles in Iran [28]. That study used MCDM methods to demonstrate that the best cycle can be obtained with little or no effort. The FAHP was noted to offer the best decision without using a lot of effort. But the given recommendations pointed out that the Fuzzy TOPSIS methods can offer better results when used in a different context such as mobile apps.

Some studies advocate the combined use of Fuzzy TOPSIS and FAHP methods [29]. But it was highlighted that this is also conditional on the need to either assign weights or ultimately rank the judgments. With little focus being given on

[1] (NCTM), N. C. (2008). Retrieved from www.nctm.org

the use of the Fuzzy TOPSIS approach to rank MMLAs, this study, therefore, deems the use of the Fuzzy TOPSIS approach is best suitable to developing a web application for ranking MMLAs.

Researchers in [30] used the fuzzy TOPSIS to assess 34 systems to locate the most adequate business intelligence for enterprise systems. This involved the computation of evaluation scores and the assigning of ranks to the systems. This approach was justified in its use citing that it allows selection, assessment, and purchasing. The findings were in line with this proposition and considerations can be made that the same can be made with regards to MMLAs, whereas the focus was primarily on quality and user-enhancing aspects of the MMLAs.

A study analyzed the use of the Fuzzy TOPSIS and AHP approaches to assign weights and rank alternatives respectively [31]. The results showed that both approaches are viable in dealing with MCDM issues. Hence, the same expectations can be individually made with regards to the Fuzzy TOPSIS approach.

MMLAs are an innovative approach and their integration in education offers a widespread number of benefits. Such benefits tend to be more when weighed against traditional learning methods. One can thus contend that aspects relating to convenience, easy access, mobility and time are major beneficial attributes of using MMLAs. However, there are also a series of problems that can undermine the use of MMLAs. These problems relate to the purpose over actual results, quality, reliability, user satisfaction, software, costs, and accessibility (internet access) aspects of MMLAs. Any challenge pertaining to these aspects can hinder the use of MMLAs. The notable idea is that the use of the Fuzzy TOPSIS which is deemed to be an optimal, viable solution to select high-quality MLAs.

In this study, the application of the Fuzzy TOPSIS approach can be based on determining the best MLAM which is either reliable, fast, easy to use, of high quality, cheaper, covers a lot of topics, etc. However, all these elements can be embodied under user experience, and hence choosing the best and high-quality MLAM that enhances user experience.

Well-known research issues have been shown to support the need for increased development of quality and user satisfaction models. These researches have also revealed that existing MLAs have not greatly impacted learning. Numerous observations which expose a mismatch between MLAMs user ratings and their contributions to improving the learning experience have further reinforced this. The quality of MLAMs has also been confirmed to be below par in most scenarios. There is a need for quality standards for the proper evaluation of mobile Mathematics learning applications.

For this purpose, ISO/IEC 25010 was established in 2011. It is a product quality standard that provides a platform where developers can evaluate and choose the software properties they wish to focus on. ISO 25010 considers the best software as that which can meet at least eight of the stated quality features.

- Functional suitability: This ensures that the laid down criteria are always met by the developed MMLAs. Sub-characteristics such as functional completeness, correctness, and appropriateness must also be met. Hence, for the MMLAs to function properly they must meet all the laid down objectives. The objectives and tasks include assigning the needed results in the proper way and with high precision. This ensures that objectives are met and tasks completed.

To extent which a product or system offers the right functionality is satisfied by some given sub-characteristics under certain conditions. These sub-characteristics must show the following:

- Functional completeness: All user objectives and specified tasks must have a degree of functionality to be satisfied. This is called functional completeness.
- Functional correctness: The product or system must also provide the right results at a high degree of correctness.
- Functional appropriateness: Specified tasks and objectives must be facilitated and accomplished to a high degree.
- Performance efficiency: The developed software or application is required to work efficiently at a rate that does not involve the consumption of many resources. This can be achieved by using a few and limited kinds of resources when functioning. To be efficient, it must provide a high degree of result at a minimum time – that is, it should not take a long time to complete the required task. Researcher in [32] has shown that an application's efficiency is shown by its ability to meet required tasks at its maximum limits. This can also be gauged by considering the performance in comparison to the number of resources under the given conditions. Performance efficiency includes:
- Time behavior: While executing its task to comply requirements, a system time behavior is measured by the extent to which the reaction and processing are put in it.
- Resource utilization: While performing, it is measured by the extent to which amounts and types of resources are put into it.
- Capacity: It refers to extent which a system's maximum limits align with the given requirements.
- Compatibility: It is the degree which a system can deliver information while performing its required functions under shared conditions with other systems. This feature includes:
- Co-existence: Sharing a same environment and resources with other products can cause harm to a product and prevent it from performing as intended. The degree to which it can prevent this is called co-existence.

- Interoperability: The use and exchange of information between two or more systems are possible. The degree to which they can do this is called interoperability.
- Usability: This refers to how well a system may be utilized by specific people to achieve specific objectives. To do this it must be effective, efficient, and satisfy the specified task. The usability includes:
- Appropriateness recognizability: Users must see if their system meets their needs. Appropriateness recognizability is the measure of how well this can be performed.
- Learnability: This refers to the extent to which a system may be used by a specific group of people. It signifies that the user uses the system to attain certain learning objectives. It provides effectiveness, efficiency, risk-free operation, and satisfaction in a certain usage environment.
- Operability: This refers to how well a system is designed to be simple to use, operate, and navigate.
- User error protection: This refers to how well a product or system protects users from making mistakes.
- User interface aesthetics: This is the extent to which a user interface allows for satisfying and enjoyable engagement.
- Accessibility: This refers to a product's or system's ability to be utilized by people with a wide range of features and abilities. It assists users in achieving a certain goal in a specific setting.
- Reliability: A system is considered reliable if it performs specified functions to a certain degree under given conditions for a specified period of time. Reliability includes:
- Maturity: This is the degree to which a system satisfies reliability requirements under regular operation.
- Availability: This is the degree to which a system is operational and available for a certain task when it is required.
- Fault tolerance: This is the degree to which a system can work even if it has defects in its hardware or software.
- Recoverability: In the event of an interruption or failure, this is the degree to which a system recovers the data directly damaged. It also brings the system back to its original state.
- Security: A system's information and data must be properly protected in order for other systems to have the right level of data access for their types and levels of authorization. This is called security. Security includes:
- Confidentiality: This is the degree to which a system's data is exclusively available to those who have been granted access.
- Integrity: Unauthorized users should not be able to access or modify a system's programs or data, thus it must be able to identify and prohibit them.
- Non-repudiation: This is the degree to which a system's actions or occurrences may be verified to have occurred. This eliminates the possibility of future repudiation for the events or actions.
- Accountability: This is the degree to which a system's actions may be traced back to another entity.
- Authenticity: This is the degree to which a subject's or resource's identification may be proven to be the one asserted.
- Maintainability: This shows how effective and efficient a system is. This is done in order to improve, fix, or adapt it to diverse environmental and other constraints. Maintainability includes:
- Modularity: This indicates how many separate components make up a system. When one component is changed, the effect on the other components is modest.
- Reusability: This is the extent to which an asset can be used in a variety of different assets. It can also be used to construct or create new items.
- Analyzability: There are times when one or more pieces of a system or product need to be changed. This is done to figure out what's wrong with the parts or what's causing them to fail. It can also be used to locate pieces that need to be fixed or modified. Analyzability refers to the degree of efficacy and efficiency with which something can be accomplished.
- Modifiability: The capacity of a system to be modified correctly without adding errors or degrading its existing quality is referred to as modifiability.
- Testability: This is the effectiveness and efficiency of a system that is used to set test criteria. It also refers to the extent to which tests can be performed to see if the requirements have been met.
- Portability: This refers to how easily a system, product, or component may be moved from one piece of hardware or software to another. It also demonstrates how quickly it may be moved from one operational or application setting to another. Portability includes:
- Adaptability: This refers to how well a system can be converted to new or changed hardware, software, or other operational or usage settings successfully and efficiently.
- Installability: This is the level of efficacy and efficiency with which a system can be successfully installed and/or uninstalled in a given environment.
- Replaceability: This is the degree to which one software can be replaced by another for the same purpose in the same environment or setting.

- Flexibility: It is the ability of the software to adapt itself easily to different user/system related requirements.
- Effectiveness: It is defined as the degree to which software performs tasks properly.

As the reviewed relevant studies have implied that despite the rapid expansion of mobile learning applications, there is a paucity of research on the most effective methods for picking a good mobile application, particularly for mobile apps for mathematics. Furthermore, only a few researches have used multi-criteria decision-making methods (MCDM) for that aim. Therefore this study aims to fill this gap in the literature.

III. Methodology

A. Triangular Fuzzy Numbers(TFN)

The transformation process of fuzzy member functions is based on the assumption or rule that an equal membership function ranging from 0.25-0.30 can be assigned to each rank (Torfi, Farahani & Rezapour, 2010). For instance, a low triangular fuzzy member of 0.000 can be assigned to a very low fuzzy variable (see Fig. 1) Table I shows the linguistic variables used for the fuzzification of the criteria. Table II shows the linguistic variables used for the fuzzification of weights.

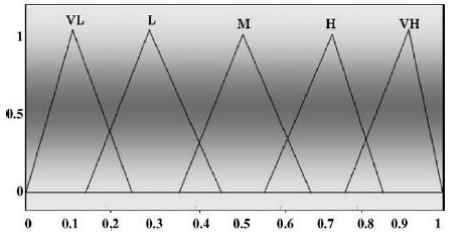


Fig. 1. Fuzzy Triangular Membership Functions.

TABLE I. Transformation for Fuzzy Membership Functions [31]

Rank	Sub-criteria grade	Membership function
Very low (VL)	1	(0.00,0.10,0.25)
Low (L)	2	(0.15,0.30,0.45)
Medium (M)	3	(0.35,0.50,0.65)
High (H)	4	(0.55,0.70,0.85)
Very high (VH)	5	(0.75,0.90,1.00

TABLE II. Linguistic Variables for the Weight

Rank	Rating	Membership function
Unnecessary (U)	1	(0,0.1,0.25)
Not Important (NI)	2	(0.15,0.30,0.45)
Important (I)	3	(0.35,0.5,0.65)
Very Important (VI)	4	(0.55,0.7,0.85)
Essential (E)	5	(0.75,0.9,1.0)

B. Mobile Mathematics Learning Applications as Alternatives

Table III shows a total of six MMLAs with user ratings of at least 4.2 and 100 000 downloads were selected from Google Play Store and Apple Store. Thus, these six MMLAs constitute a sample of MMLAs that were used in this study to create a platform upon which the best mobile mathematics applications in terms of high-quality and user enhancing experience can be selected.

TABLE III. Alternatives

Math application	Google Store		Apple Store	
	User ratings	Downloads in 2018	User ratings	Downloads in 2018
yHomework Math Solver	4.2	1 000 000+	4.6	3 000 000+
Cymath	4.5	100 000+	4.3	100 000+
Malmath	4.6	500 000+	N/A	N/A
Math 42	4.6	500 000+	4.5	3 400 000+
MathPapa	4.7	500 000+	4.7	500 000+
PhotoMath	4.7	50 000 000+	4.8	100 000 000+

1) yHomework - math solver: Math Solver in Fig. 2 specifically focues on dealing with algebra issues but also incorporates mathematicals topics involving the use of graphs, solving inequalities and other types of equations. The applications simply requires users to enter an equation and it automatically computes the answer for the user.

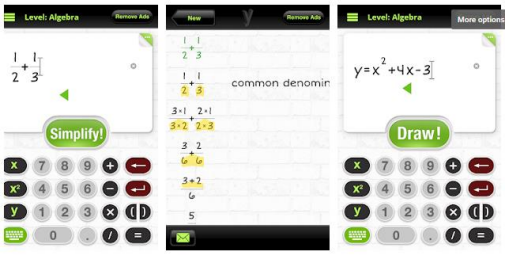


Fig. 2. Screenshot of yHomework - Math Solver.

2) Cymath: Cymath depicted in Fig. 3 solves math problems such as algebra (eg. quadratic equations, complex numbvers, exponents, logarithyms factoring etc.) and calculus (eg. trigonometric substitution, integration, u-substitution, chain rule etc.) using the same mathematical engines. As such, it simply allows users to enter the mathematical problem and then automatically computes the answer for them.

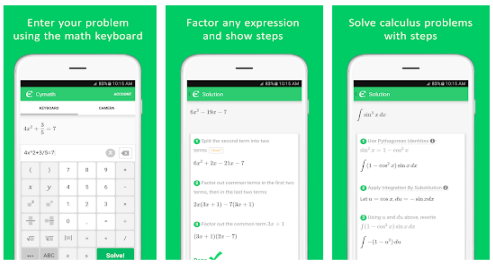


Fig. 3. Screenshot of CyMath.

3) Malmath: MalMath shown in Fig. 4 is used to solve mathematical problems with graphical interface and instructions. It is freely available working online and offline together. It helps in dealing with topics involving the solving of algebra, integrals, equations, derivatives, trigonometry, logarithms, limits, etc. It provides solving process as well and inteded for high school and university students and instructors/teachers.

Fig. 4. Screenshot of MalMath.

4) Math42: Math42 given in Fig. 5 provides innovative, step by step checking process guide to solving problems and it also includes test center. It also includes features such as autocomplete formula entry recommendations during problem solving.

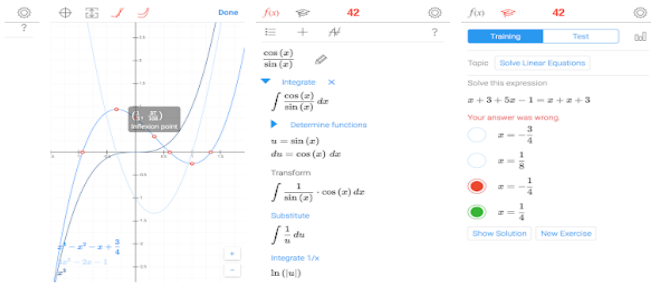


Fig. 5. Screenshot of Math42.

5) MathPapa: MathPapa depicted in Fig. 6 provides aid in solving particularly linear equations and quadratic equations and inequalities, graphs.

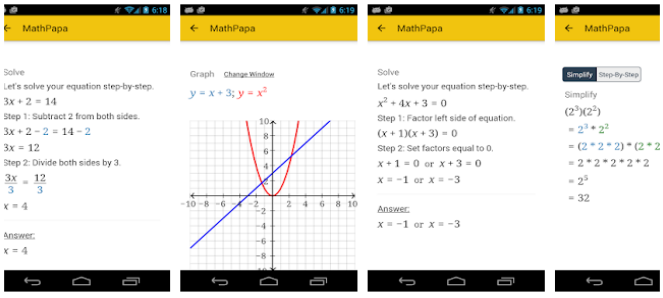


Fig. 6. Screenshot of MathPapa.

6) PhotoMath: PhotoMath shown in Fig. 7 provides guide for solving mathematical problems. It includes monitoring for assignments and exams. Photomath is freely available and works online. It has scanned text and handwriting text recognition feature.

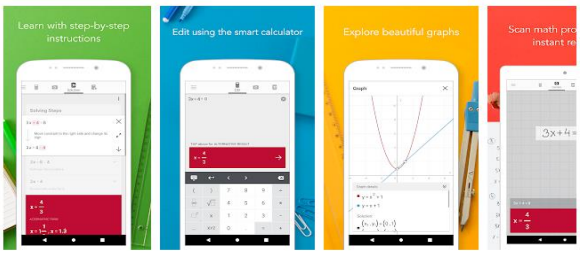


Fig. 7. Screenshot of PhotoMath.

C. Fuzzy Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) Method

The steps taking in the research include:

Step 1: A survey was set and given to the decision makers. The decision makers then evaluated the questions for six alternative set for the selected criteria.

Let m be number of alternatives {A1, A2 …, Am} (m ≥ 2)

For this research, m=6.

Let n be the Number of Criteria {C1, C2… Cn} (n ≥ 2)

For this research, n=12.

Let w be the vector of Criteria Weights (0 ≤ w ≤ 1), while $\Sigma_1^n w_i = 1$.

Let DM be the number of Decision Makers that assess the alternatives (A) and all the Criteria (C) while {DM1, DM2, …, DMK} (K ≥ 2)

For the research, DM=2.

Step 2: The results from Step 1 were imputed into the web based software for the matrix which thereafter, went through the process of calculating the normalized fuzzy decision.

The Decision Making Matrix

$$X^k = \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} \begin{bmatrix} x_{11}^k & x_{12}^k & \cdots & x_{1n}^k \\ x_{21}^k & x_{22}^k & \cdots & x_{2n}^k \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}^k & x_{m2}^k & \cdots & x_{mn}^k \end{bmatrix} \quad (1)$$

$$C_1 \quad C_2 \quad \ldots \quad C_n$$

The normalized fuzzy decision matrix

$$Y^k = \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} \begin{bmatrix} x_{11}^k & x_{12}^k & \cdots & x_{1n}^k \\ x_{21}^k & x_{22}^k & \cdots & x_{2n}^k \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}^k & x_{m2}^k & \cdots & x_{mn}^k \end{bmatrix} \quad (2)$$

$$C_1 \quad C_2 \quad \ldots \quad C_n$$

Where

$$y_{ij}^k = \begin{cases} \left(\frac{a_{x_{ij}^k}}{max_i d_{x_{ij}^k}}, \frac{b_{x_{ij}^k}}{max_i d_{x_{ij}^k}}, \frac{c_{x_{ij}^k}}{max_i d_{x_{ij}^k}} \right) if\ j \in B \\ \left(\frac{min_i a_{x_{ij}^k}}{a_{x_{ij}^k}}, \frac{min_i a_{x_{ij}^k}}{b_{x_{ij}^k}}, \frac{min_i a_{x_{ij}^k}}{c_{x_{ij}^k}} \right) if\ j \in C \end{cases} \quad (3)$$

Step 3: Thereafter, the result was imputed into another matrix to obtain the weighted normalized fuzzy decision.

The weighted normalized fuzzy decision matrix

$$v^k = \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} \begin{bmatrix} v_{11}^k & v_{12}^k & \cdots & v_{1n}^k \\ v_{21}^k & v_{22}^k & \cdots & v_{2n}^k \\ \vdots & \vdots & \ddots & \vdots \\ v_{m1}^k & v_{m2}^k & \cdots & v_{mn}^k \end{bmatrix} \quad (4)$$

$C_1 \quad C_2 \quad \ldots \quad C_n$

Where $v_{ij}^k = w_j y_{ij}^k = (w_j a_{ij}^k, w_j b_{ij}^k, w_j c_{ij}^k)$

The matrices v^k form the basis of the f weighted normalized fuzzy decision matrices for each alternative A_i

Step 4: The next step was to determine the positive idea solution, A^+ and the negative idea solution, A^- using the following equations:

The positive ideal solution A^+

$$\text{Where } A^+ = \begin{matrix} DM_1 \\ DM_2 \\ \vdots \\ DM_m \end{matrix} \begin{bmatrix} v_1^{1+} & v_2^{1+} & \cdots & v_n^{1+} \\ v_1^{2+} & v_2^{2+} & \cdots & v_n^{2+} \\ \vdots & \vdots & \ddots & \vdots \\ v_1^{k+} & v_2^{k+} & \cdots & v_n^{k+} \end{bmatrix} \quad (5)$$

$C_1 \quad C_2 \quad \ldots \quad C_n$

While $v_j^{k+} = max_i v_{ij}^k$

The negative ideal solution A^-

$$\text{Where } A^- = \begin{matrix} DM_1 \\ DM_2 \\ \vdots \\ DM_m \end{matrix} \begin{bmatrix} v_1^{1-} & v_2^{1-} & \cdots & v_n^{1-} \\ v_1^{2-} & v_2^{2-} & \cdots & v_n^{2-} \\ \vdots & \vdots & \ddots & \vdots \\ v_1^{k-} & v_2^{k-} & \cdots & v_n^{k-} \end{bmatrix} \quad (6)$$

$C_1 \quad C_2 \quad \ldots \quad C_n$

While $v_j^{k-} = min_i v_{ij}^k$

Step 5: Finally, the system calculated the relative ideal situation of each alternatives. Then it ranked the alternatives accordingly starting from the closest to the ideal situation. Then the results were obtained.

The distances of each alternative A_i represented by matrix W_i from positive ideal solution (PIS);

$$d_i^+ = \Sigma_1^k \Sigma_1^n \, d(v_{ij}^k v_j^{k+}) \quad (7)$$

The distances of each alternative A_i represented by matrix W_i from negative ideal solution (NIS);

$$d_{\bar{i}} = \Sigma_1^k \Sigma_1^n \, d(v_{ij}^k v_j^{k-}) \quad (8)$$

Using these distances, the relative closeness coefficients RC_i to PIS

$$\text{Where } RC_i = \frac{d_i^-}{d_i^- + d_i^+} \quad (9)$$

According to the descending values of RC_i, all alternatives A_i are rank ordered and the best one is selected.

D. Evaluation Criteria Framework and Ranking

The MCDM evaluation criteria were based upon the two aspects that were adopted from ISO/IEC 25010. The criteria are; Functional completeness (C_1), Functional correctness (C_2), Functional appropriateness (C_3), Resource utilization (C_4), Time behavior (C_5), Appropriateness recognizability (C_6), Learnability (C_7) . Confidentiality (C_8), Effectiveness (C_9), Efficiency (C_{10}), Flexibility (C_{11}), Satisfaction in Usefulness (C_{12}). The alternatives are; yHomework Math Solver(A_1),Cymath(A_2), Malmath(A_3), Math 42(A_4),MathPapa(A_5), PhotoMath(A_6).

Two decision-makers were involved in the evaluation process of the alternatives. The first expert (DM_1) has a background in Educational technology. The second expert has a background in computer information systems (DM_2). The rationale for involving the limited number of decision-makers lies in the challenge of locating decision-makers with the proper area of expertise who not only have knowledge and experience on software quality standards but also have an adequate background on how to evaluate the software.

IV. Results

The ranking process starts with the two decision makers evaluate six alternatives by using the twelve criteria derived from ISO/IEC 25010 software quality standard metrics. The linguistic scale given in Table I was used to evaluate criteria by the experts. The two decision matrices of the evaluated alternatives were given in Table IV and Table V respectively. Later, the evaluation of the decision makers was converted into fuzzy scales. Fuzzy decision matrices for decision maker 1 and decision maker 2 were shown in Table VI and Table VII, respectively. By using the linguistic weights given in Table II, the weighted decision matrix was calculated and is specified in Table VIII. The weighted decision matrix is then normalized and ideal solutions were calculated which were given in Table IX. The normalized positive and negative ideal solution matrices are shown in Table X and Table XI respectively. The Table XII shows the closeness to the ideal solutions from highest to lowest and the final ranking of the alternatives. The results revealed that PhotoMath(A_6) > Malmath (A_3)> Math 42(A_4)> Cymath(A_2)> MathPapa(A_5)> yHomework Math Solver (A_1) where PhotoMath has the highest rank whereas yHomework Math Solver has the lowest rank in terms of selected criteria according to fuzzy TOPSIS ranking procedure.

To authors' knowledge, studies that employ MCDM techniques to evaluate the quality of mobile apps particularly for Mathematics are quite limited. This constitutes the essential driving motivation to conduct such research. The study has some superior features as compared to the earlier studies in the literature. The authors of an earlier study in [6] only considered five alternatives and merely one decision maker whereas this research included six alternatives and two decision makers. Another study applied ELECTRE I to five alternatives with only one decision maker [11] whereas the number of decision makers in this study is two and the number of alternatives are more. It was inferred that fuzzy TOPSIS method can be quite effortlessly employed. The fuzzy TOPSIS procedure is a popular technique used in other studies where

researchers used fuzzy TOPSIS methods to evaluate four general learning applications with 175 students using 25 criteria [33]. Earlier relevant studies have integrated FAHP and conventional TOPSIS techniques [6] or used TOPSIS to evaluate 6 language learning apps with six experts and 17 criteria [34], [35] whereas in the absence of precise performance ratings fuzzy TOPSIS is the prominent technique over conventional TOPSIS which justifies the use of fuzzy TOPSIS in this study.

TABLE IV. DM_1 DECISION MATRIX

DM_1	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**	**C_8**	**C_9**	**C_{10}**	**C_{11}**	**C_{12}**
A_1	H	H	L	VH	M	L	VH	L	H	M	H	VH
A_2	L	VL	VH	H	VH	VL	H	L	M	VH	M	H
A_3	H	M	VH	L	VL	H	VL	M	VH	M	VH	L
A_4	VL	VH	M	VH	H	H	M	VH	H	VL	M	M
A_5	VH	H	VL	M	VH	M	VH	H	VL	VH	VL	VH
A_6	H	VL	H	VL	M	VH	H	VL	L	L	VL	H

TABLE V. DM_2 DECISION MATRIX

DM_2	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**	**C_8**	**C_9**	**C_{10}**	**C_{11}**	**C_{12}**
A_1	VH	M	L	L	VH	H	H	L	M	VH	H	H
A_2	H	VH	L	VL	H	VL	L	VH	VH	H	M	M
A_3	L	M	M	H	L	M	H	VH	VL	VL	VH	VH
A_4	M	VL	VH	H	VH	VH	VL	M	H	M	H	M
A_5	VH	VH	H	M	M	H	VH	VL	VH	VH	VL	VL
A_6	H	L	VL	VH	VL	VL	H	H	M	H	L	VL

TABLE VI. FUZZY DM_1 DECISION MATRIX

W	E	VI	NI	E	I	NI	E	I	I	E	VI	E
D M_1	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**	**C_8**	**C_9**	**C_{10}**	**C_{11}**	**C_{12}**
A_1	(0.65,0.8, 0.95)	(0.65,0.8, 0.95)	(0,0.1,0.2)	(0.9,1,**1)**	(0.35,0.5, 0.65)	(0,0.1,0.2)	(0.9,1,1)	(0,0.1,0.2)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0.9,1,1)
A_2	(0,0.1,0.2)	(0,0,0.1)	(0.9,1,1)	(0.65,0.8, 0.95)	(0.9,1,1)	(0,0,0.1)	(0.65,0.8, 0.95)	(0,0.1,0.2)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)
A_3	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.9,1,1)	(0,0.1,0.2)	(0,0,0.1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.35,0.5, 0.65)	(0.9,1,1)	(0,0.1,0.2)
A_4	(0,0,0.1)	(0.9,1,1)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.65,0.8, 0.95)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0.35,0.5, 0.65)	(0.35,0.5, .65)
A_5	(0.9,1,1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0.9,1,1)	(0,0,0.1)	(0.9,1,1)
A_6	(0.65,0.8, 0.95)	(0,0,0.1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0,0.1,0.2)	(0,0.1,0.2)	(0,0,0.1)	(0.65,0.8, 0.95)
w	(0.75,0.9, 1)	(0.55,0.7, 0.85)	(0.15,0.3, 0.45)	(0.75,0.9, 1)	(0.35,0.55 ,0.65)	(0.15,0.3, 0.45)	(0.75,0.9, 1)	(0.35,0.55 ,0.65)	(0.35,0.55 ,0.65)	(0.75,0.9, 1)	(0.55,0.7, 0.85)	(0.75,0.9, 1)

TABLE VII. FUZZY DM2 DECISION MATRIX

W	E	VI	NI	E	I	NI	E	I	I	E	VI	E
DM$_2$	**C$_1$**	**C$_2$**	**C$_3$**	**C$_4$**	**C$_5$**	**C$_6$**	**C$_7$**	**C$_8$**	**C$_9$**	**C$_{10}$**	**C$_{11}$**	**C$_{12}$**
A$_1$	(0.9,1,1)	(0.35,0.5, 0.65)	(0,0.1,0.2)	(0,0.1,0.2)	(0.9,1,**1)**	(0.65,0.8, 0.95)	(0.65,0.8, 0.95)	(0,0.1,0.2)	(0.35,0.5, 0.65)	(0.9,1,1)	(0.65,0.8, 0.95)	(0.65,0.8, 0.95)
A$_2$	(0.65,0.8, 0.95)	(0.9,1,1)	(0,0.1,0.2)	(0,0,0.1)	(0.65,0.8, 0.95)	(0,0,0.1)	(0,0.1,0.2)	(0.9,1,1)	(0.9,1,1)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.35,0.5, 0.65)
A$_3$	(0,0.1,0.2)	(0.35,0.5, 0.65)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0,0.1,0.2)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0.9,1,1)	(0,0,0.1)	(0,0,0.1)	(0.9,1,1)	(0.9,1,1)
A$_4$	(0.35,0.5, 0.65)	(0,0,0.1)	(0.9,1,1)	(0.65,0.8, 0.95)	(0.9,1,1)	(0.9,1,1)	(0,0,0.1)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)
A$_5$	(0.9,1,1)	(0.9,1,1)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0.9,1,1)	(0,0,0.1)	(0.9,1,1)	(0.9,1,1)	(0,0,0.1)	(0,0,0.1)
A$_6$	(0.65,0.8, 0.95)	(0,0.1,0.2)	(0,0,0.1)	(0.9,1,1)	(0,0,0.1)	(0,0,0.1)	(0.65,0.8, 0.95)	(0.65,0.8, 0.95)	(0.35,0.5, 0.65)	(0.65,0.8, 0.95)	(0,0.1,0.2)	(0,0,0.1)

TABLE VIII. WEIGHTED DECISION MATRIX

DM	**C$_1$**	**C$_2$**	**C$_3$**	**C$_4$**	**C$_5$**	**C$_6$**	**C$_7$**	**C$_8$**	**C$_9$**	**C$_{10}$**	**C$_{11}$**	**C$_{12}$**
A$_1$	(0.65,0.9, 1)	(0.35,0.65, 0.95)	(0,0.1,0. 2)	(0,0.55,1)	(0.35,0. 75,1)	(0,0.45,0.9 5)	(0.65,0.9, 1)	(0,0.1,0. 2)	(0.35,0.65, 0.95)	(0.35,0. 75,1)	(0.65,0.8,0 .95)	(0.65,0.9,1)
A$_2$	(0,0.45,0. 95)	(0,0.5,1)	(0,0.55, 1)	(0,0.4,0.9 5)	(0.65,0. 9,1)	(0,0,0.1)	(0,0.45,0. 95)	(0,0.55, 1)	(0.35,0.75, 1)	(0.65,0. 9,1)	(0.35,0.5,0 .65)	(0.35,0.65, 0.95)
A$_3$	(0,0.45,0. 95)	(0.35,0.5,0 .65)	(0.35,0. 75,1)	(0,0.45,0. 95)	(0,0.05, 0.2)	(0.35,0.65, 0.95)	(0,0.4,0.9 5)	(0.35,0. 75,1)	(0,0.5,1)	(0,0.25, 0.65)	(0.9,1,1)	(0,0.55,1)
A$_4$	(0,0.25,0. 65)	(0,0.5,1)	(0.35,0. 75,1)	(0.65,0.9, 1)	(0.65,0. 9,1)	(0.65,0.9,1)	(0,0.25,0. 65)	(0.35,0. 75,1)	(0.65,0.8,0 .95)	(0,0.25, 0.65)	(0.35,0.65, 0.95)	(0.35,0.5,0 .65)
A$_5$	(0.9,1,1)	(0.65,0.9,1)	(0,0.4,0. 95)	(0.35,0.5, 0.65)	(0.35,0. 75,1)	(0.35,0.65, 0.95)	(0.9,1,1)	(0,0.4,0. 95)	(0,0.5,1)	(0.9,1,1)	(0,0,0.1)	(0,0.5,1)
A$_6$	(0.65,0.8, 0.95)	(0,0.05,0.2)	(0,0.4,0. 95)	(0,0.5,1)	(0,0.25, 0.65)	(0,0.5,1)	(0.65,0.8, 0.95)	(0,0.4,0. 95)	(0,0.3,0.65)	(0,0.45, 0.95)	(0,0.05,0.2)	(0,0.4,0.95)

TABLE IX. NORMALIZED WEIGHTED DECISION MATRIX WITH IDEAL SOLUTIONS

Ideal	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
A_1	(0.4875, 0.81,19	(0.1925,0.455,0.8075)	(0,0.03,0.09)	(0,0.495,1)	(0.1225,0.4125,0.65)	(0,0.135,0.4275)	(0.4875, 0.81,1)	(0,0.055,0.13)	(0.1225,0.3575,0.6175)	(0.2625,0.675,1)	(0.3575,0.56,0.8075)	(0.4875,0.81,1)
A_2	(0,0.405, 0.95)	(0,0.35,0.85)	(0,0.165, 0.45)	(0,0.36,0.95)	(0.2275,0.495,0.65)	(0,0,0.045)	(0,0.405, 0.95)	(0,0.302,0.65)	(0.1225,0.4125,0.65)	(0.4875,0.81,1)	(0.1925,0.35,0.5525)	(0.2625,0.585,0.95)
A_3	(0,0.405, 0.95)	(0.1925,0.35,0.5525)	(0.0525,0.225,0.45)	(0,0.405, 0.95)	(0,0.0275, 0.13)	(0.0525,0.195,0.4275)	(0,0.36,0.95)	(0.1225,0.4125,0.65)	(0,0.275,0.65)	(0,0.225,0.65)	(0.495,0.7, 0.85)	(0,0.495, 1)
A_4	(0,0.225, 0.65)	(0,0.35,0.85)	(0.0525,0.225,0.45)	(0.4875, 0.81,1)	(0.2275,0.495,0.65)	(0.0975,0.27,0.45)	(0,0.225, 0.65)	(0.1225,0.4125,0.65)	(0.2275,0.44,0.6175)	(0,0.225,0.65)	(0.1925,0.455,0.8075)	(0.2625,0.45,0.65)
A_5	(0.675,0.9,1)	(0.3575,0.63,0.85)	(0,0.12,0.4275)	(0.2625, 0.45,0.65)	(0.1225,0.4125,0.65)	(0.0525,0.195,0.4275)	(0.675,0.9,1)	(0,0.22,0.6175)	(0,0.275,0.65)	(0.675, 0.9,1)	(0,0,0.085)	(0,0.45,1)
A_6	(0.4875, 0.72,0.95)	(0,0.035,0.17)	(0,0.12,0.4275)	(0,0.45,1)	(0,0.1375, 0.4225)	(0,0.15,0.45)	(0.4875, 0.72,0.95)	(0,0.22,0.6175)	(0,0.165,0.4225)	(0,0.405,0.95)	(0,0.035,0.17)	(0,0.36,0.95)
A^+	(0.675,0.9,1)	(0.3575,0.65,0.85)	(0.0525,0.225,0.45)	(0.4875, 0.81,1)	(0.2275,0.495,0.65)	(0.0525,0.195,0.45)	(0.675,0.9,1)	(0.1225,0.4125,0.65)	(0.1225,0.4125,0.65)	(0.675, 0.9,1)	(0.495,0.7, 0.85)	(0.4875,0.81,1)
A^-	(0,0.225, 0.65)	(0,0.035,0.17)	(0,0.03,0.09)	(0,0.36,0.95)	(0,0.0275, 0.13)	(0,0.135,0.4275)	(0,0.225, 0.65)	(0,0.055,0.13)	(0,0.165,0.4225)	(0,0.225,0.65)	(0,00.085)	(0,0.36,0.95)

TABLE X. NORMALIZED POSITIVE IDEAL SOLUTION MATRIX

FPIS	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	d+i
A_1	0.14077908	0.22114758	0.41054080	0.42242603	0.10237798	0.07088723	0.14077908	0.63498688	0.06388466	0.32763356	0.16645820	0.00000000	2.70190108
A_2	0.34985711	0.36414569	0.06722165	0.53312170	0.00000000	0.45052053	0.63198101	0.13077493	0.00000000	0.14077908	0.49143497	0.26457513	3.42441182
A_3	0.63198101	0.43310651	0.00000000	0.49572548	0.71148319	0.02250000	0.66781360	0.00000000	0.15462320	0.85440037	0.00000000	0.42242603	4.39405939
A_4	0.85440037	0.36414569	0.00000000	0.00000000	0.00000000	0.07937254	0.85440037	0.00000000	0.07407766	0.85440037	0.30386400	0.51862800	3.90328902
A_5	0.00000000	0.02000000	0.11157957	0.51862800	0.10237798	0.02250000	0.00000000	0.20764051	0.15462320	0.00000000	1.07559286	0.45696690	2.66990902
A_6	0.21591376	0.93980162	0.11157957	0.45696690	0.44363790	0.05425634	0.21591376	0.20764051	0.34353251	0.63198101	0.99312638	0.53312170	5.14747195

TABLE XI. NORMALIZED NEGATIVE IDEAL SOLUTION MATRIX

FINS	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	d-i
A_1	0.73752 5423	0.77146 5056	0.00000 0000	0.14396 1800	0.65086 6410	0.00000 0000	0.73752 5423	0.00000 0000	0.28298 9988	0.58988 8761	0.93712 7704	0.53312 1703	5.38447 2269
A_2	0.34985 7114	0.74941 6440	0.38448 0169	0.00000 0000	0.71148 3193	0.40562 4518	0.34985 7114	0.57589 6041	0.34353 2507	0.73752 5423	0.59448 1567	0.27128 1680	5.47343 5765
A_3	0.34985 7114	0.07366 3180	0.41054 0802	1.03495 2088	0.00000 0000	0.06722 1648	0.32897 5683	0.63498 6877	0.25269 7942	0.00000 0000	1.07559 2860	0.14396 1800	4.37244 9993
A_4	0.00000 0000	0.74941 6440	0.40367 0638	0.53312 1703	0.71148 3193	0.14798 6486	0.00000 0000	0.63498 6877	0.36180 3929	0.00000 0000	0.86103 6197	0.34794 9350	4.75145 4812
A_5	0.56235 2490	0.60924 8723	0.34929 3931	0.34794 9350	0.65086 6410	0.06722 1648	0.85440 0375	0.51466 6154	0.25269 7942	0.85440 0375	0.00000 0000	0.10295 6301	5.16605 3698
A_6	0.64361 7705	0.00000 0000	0.34929 3931	0.10295 6301	0.31250 0000	0.02704 1635	0.64361 7705	0.51466 6154	0.00000 0000	0.34985 7114	0.09192 3882	0.00000 0000	3.03547 4426

TABLE XII. FINAL RANKING OF ALTERNATIVES

Rank	d-i	d+i	D+i+D-i	Cci	Rank
yHomework Math Solver (A_1)	2.701901083	5.384472269	8.086373351	0.334130143	6
Cymath(A_2)	3.42441182	5.473435765	8.897847585	0.384858449	4
Malmath (A_3)	4.394059393	4.372449993	8.766509386	0.501232497	2
Math 42(A_4)	3.903289017	4.751454812	8.654743829	0.450999948	3
MathPapa(A_5)	2.669909017	5.166053698	7.835962715	0.340725079	5
PhotoMath(A_6)	5.147471948	3.035474426	8.182946373	0.629048721	1

V. CONCLUSION

The user ratings of mobile applications found at the app stores can sometimes be not sufficient for revealing the essential quality of the mobile applications. Therefore precise and easy techniques are desired. By developing a multi criteria decision making evaluation to rank mobile mathematics learning applications, this study intends to enhance the use and effectiveness of mobile Mathematics learning applications thereby improving the quality of learning across all learning platforms. In addition, the study also helps in setting standards in relation to ISO practices upon which the quality and usability of MMLAs can be determined. It also contributes to the research of mobile applications through the use of fuzzy TOPSIS.

VI. FUTURE WORK

Tends to focus towards increasing the number of decision makers, alternatives as well as comparing and contrasting the efficiency of different MCDM methods applied to improve the precision of the selection process. It is recommended that user friendly interface or software could be initiated for the service of decision makers.

REFERENCES

[1] S. Başaran, "Multi-Criteria Decision Analysis Approaches for Selecting and Evaluating Digital Learning Objects," Procedia Computer Science, vol. 102, pp. 251–258, 2016. DOI: https://doi.org/10.1016/j.procs.2016.09.398.

[2] M. L. Crescente and D. Lee, "Critical issues of m-learning: design models, adoption processes, and future trends," Journal of the Chinese Institute of Industrial Engineers, vol. 28, no. 2, pp. 111–123, Mar. 2011.

[3] A. Drigas and M. Pappas, "A Review of Mobile Learning Applications for Mathematics," International Journal of Interactive Mobile Technologies (iJIM), vol. 9, no. 3, p. 18, Jul. 2015.

[4] R. Pierce, K. Stacey, and A. Barkatsas, "A scale for monitoring students' attitudes to learning mathematics with technology," Computers & Education, vol. 48, no. 2, pp. 285–300, Feb. 2007.

[5] M., Bjerede, K. Atkins, and C., Dede, "A Special Report: Ubiquitous Mobile Technologies and the Transformation of Schooling," Educational Technology, pp.3-7,2010.

[6] S. Başaran and Y. Haruna, "Integrating FAHP and TOPSIS to evaluate mobile learning applications for mathematics," Procedia Computer Science, vol. 120, pp. 91–98, 2017. DOI: https://doi.org/10.1016/j.procs.2017.11.214.

[7] R. H. Kay and L. Knaack, "Evaluating the learning in learning objects," Open Learning: The Journal of Open, Distance and e-Learning, vol. 22, no. 1, pp. 5–28, Feb. 2007.

[8] G. Büyüközkan and S. Güleryüz, "Multi Criteria Group Decision Making Approach for Smart Phone Selection Using Intuitionistic Fuzzy TOPSIS," International Journal of Computational Intelligence Systems, vol. 9, no. 4, p. 709, 2016.

[9] R. Trestian, A.-N. Moldovan, C. H. Muntean, O. Ormond, and G.-M. Muntean, "Quality Utility modelling for multimedia applications for Android Mobile devices," IEEE international Symposium on Broadband Multimedia Systems and Broadcasting, pp. 1-6, Jun. 2012.

[10] T.L., Leacock, and J.C., Nesbit,. "A framework for evaluating the quality of multimedia learning resources," Journal of Educational Technology & Society, 10(2), pp.44-59,2007.

[11] S. Basaran and O. J., Aduradola, "A Multi-Criteria Decision Making to Rank Android based Mobile Applications for Mathematics,"

International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, 2018. DOI: 10.14569/IJACSA.2018.090714.

[12] E. K. Zavadskas, Z. Turskis, and S. Kildienė, "State Of Art Surveys Of Overviews On Mcdm/Madm Methods," Technological and Economic Development of Economy, vol. 20, no. 1, pp. 165–179, Mar. 2014.

[13] A. Daghouri, K. Mansouri, and M. Qbadou, "Multi Criteria Decision Making methods for Information System Selection: A Comparative Study," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Dec. 2018.

[14] I. B. Huang, J. Keisler, and I. Linkov, "Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends," Science of The Total Environment, vol. 409, no. 19, pp. 3578–3594, Sep. 2011.

[15] S. Rouhani, M. Ghazanfari, and M. Jafari, "Evaluation model of business intelligence for enterprise systems using fuzzy TOPSIS," Expert Systems with Applications, vol. 39, no. 3, pp. 3764–3771, Feb. 2012.

[16] R.M., Palloff, K. Pratt, and D., Stockley, "Building learning communities in cyberspace: Effective strategies for the online classroom," The Canadian Journal of Higher Education, 31(3), p.175-178,2001.

[17] M. Virvou and E. Alepis, "Mobile educational features in authoring tools for personalised tutoring," Computers & Education, vol. 44, no. 1, pp. 53–68, Jan. 2005.

[18] L. F. Motiwalla, "Mobile learning: A framework and evaluation," Computers & Education, vol. 49, no. 3, pp. 581–596, Nov. 2007.

[19] F., Lehner, & H., Nosekabel, "The role of mobile devices in E-Learning first experiences with a wireless E-Learning environment," IEEE International Workshop on Wireless and Mobile Technologies in Education, pp. 103-106, IEEE, 2002.

[20] H. U., Hoppe, S., Eimler, & L., Bollen, "The use of mobile computing to support SMS dialogues and classroom discussions in a literature course," IEEE International Conference on Advanced Learning Technologies, pp. 550-554, IEEE,2004.

[21] M., Skillen, "Mobile Learning: Impacts on Mathematics Education.," the 20th Asian Technology Conference in Mathematics, pp. 205-214, China., 2015.

[22] G. Botzer, and M., Yerushalmy, "Mobile application for mobile learning," In Proceedings of IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2007), pp. 7-9, December 2007.

[23] N., Roberts, G., Spencer-Smith, R. Vanska, and S., Eskelinen, "From challenging assumptions to measuring effect: Researching the Nokia mobile mathematics service in South Africa," South African Journal of Education, vol. 35, no. 2, pp.1045-1045, 2015.

[24] C. Hoyles and J.-B. Lagrange, Eds., "Mathematics Education and Technology-Rethinking the Terrain," New ICMI Study Series, 2010.

[25] K., Melhuish and G., Falloon, "Looking to the future: M-learning with the iPad," Computers in New Zealand Schools: Learning, Leading, Technology, vol. 22, no.3, pp. 1–16, 2010.

[26] P., Drijvers, "Digital technology in mathematics education: Why it works (or doesn't)." In Selected regular lectures from the 12th international congress on mathematical education, pp. 135-151. Springer, Cham, 2015.

[27] M. Z. Naghadehi, R. Mikaeil, and M. Ataei, "The application of fuzzy analytic hierarchy process (FAHP) approach to selection of optimum underground mining method for Jajarm Bauxite Mine, Iran," Expert Systems with Applications, vol. 36, no. 4, pp. 8218–8226, May 2009.

[28] M. Vafaeipour, S. Hashemkhani Zolfani, M. H. Morshed Varzandeh, A. Derakhti, and M. Keshavarz Eshkalag, "Assessment of regions priority for implementation of solar projects in Iran: New application of a hybrid multi-criteria decision making approach," Energy Conversion and Management, vol. 86, pp. 653–663, Oct. 2014.

[29] S. Ballı and S. Korukoğlu, "Operating System Selection Using Fuzzy AHP and TOPSIS Methods," Mathematical and Computational Applications, vol. 14, no. 2, pp. 119–130, Aug. 2009.

[30] S. Rouhani, M. Ghazanfari, and M. Jafari, "Evaluation model of business intelligence for enterprise systems using fuzzy TOPSIS," Expert Systems with Applications, vol. 39, no. 3, pp. 3764–3771, Feb. 2012.

[31] F. Torfi, R. Z. Farahani, and S. Rezapour, "Fuzzy AHP to determine the relative weights of evaluation criteria and Fuzzy TOPSIS to rank the alternatives," Applied Soft Computing, vol. 10, no. 2, pp. 520–528, Mar. 2010.

[32] A. A. Economides, "Requirements of Mobile Learning Applications," International Journal of Innovation and Learning, vol. 5, no. 5, p. 457, 2008.

[33] K. Singh, N. Naicker and M. Rajkoomar. Selection of Learning Apps to Promote Critical Thinking in Programming Students using Fuzzy TOPSIS. International Journal of Advanced Computer Science and Applications, Vol. 12, no.10, 2021 10.14569/IJACSA.2021.0121042

[34] N. K. Ibrahim et al., "Multi-Criteria Evaluation and Benchmarking for Young Learners' English Language Mobile Applications in Terms of LSRW Skills," IEEE Access, vol. 7, pp. 146620-146651, 2019. DOI: 10.1109/ACCESS.2019.2941640.

[35] Chunhe Zhao, Balaanand Muthu, and P. Mohamed Shakeel,"Multi-Objective Heuristic Decision Making and Benchmarking for Mobile Applications in English Language Learning," ACM Transactions on Asian and Low-Resource Language Information Processing. vol. 20, no. 5, pp. 1-16, 2021. DOI:https://doi.org/10.1145/3439799.

An Optimal Execution of Composite Service in Decentralized Environment

Yashwant Dongre, Rajesh Ingle
Department of Computer Engineering, Pune Institute of Computer Technology
Savitribai Phule Pune University, Pune, India

***Abstract*—It is important for service-oriented architectures to consider about how the composition of web services affects business processes. For instance, a single web service may not have been adequate for most complex business operations, needing the use of multiple web services. This paper proposed a novel technique for optimal partitioning and execution of the services using a decentralized environment. The proposed technique is designed and developed using a genetic algorithm with multiple high task allocations on a single server. We compared three existing techniques, including meta-heuristic genetic algorithm, heuristics like Pooling-and-Greedy-Merge (PGM) technique, and Merge-by-Define-Use (MDU) technique, to a simulation of Business Process Execution Language (BPEL) partition using genetic algorithm through multiple high tasks allocation to single server node. The proposed technique is practical and advantageous. In terms of execution time, number of server requests, and throughput, the proposed technique outperformed the existing GA, PGM, and MDU techniques.**

***Keywords*—Genetic algorithm; service composition; decentralized execution; composite service**

I. Introduction

In Service-Oriented Architecture (SOA), web services are the most important and widely implemented technologies which are interoperable machines-to-machines interactions that happen over networks [1]. A Large number of connected heterogeneous devices containing objects are expected to deploy than existing deployed devices in the coming few years. These devices require a reliable connection between them anywhere and forever which provides the ability to collect data. The large availability of devices is advantageous. Traditional services such as traffic control and healthcare are experiencing a shift to a new category of service industry demands.

Decentralized execution environments for Business Process Execution Language (BPEL) processes are necessary due to numerous reasons, starting from the outsourcing of process fragments to the need for runtime performance optimizations without modifying process models [2]. Many factors make affect partitioning or finding an appropriate distribution of the business program or process. This paper focuses on these factors in order to describe them and to offer a high-level outline of a possible process partitioning approach [3]. Whenever multiple software components and service providers are intricate in business programs, a composition of web service is essential to create a composite web service that combines multiple web-based services to collaborate with one another [4]. As the scale and number of web services have increased, many service providers are offering candidate web services that are operationally corresponding but ensure dissimilar levels of non-functional parameter values [5]. Apart from functionalities, as well as non-functional necessities often defined by QoS are employed by web services to come across operator demands [5, 6], and to satisfy business needs, and service level agreements (SLAs) are becoming an integral part of web based composition process of services.

The selection of QoS-responsive web-based services is the process used to choose web services from amongst a set of candidate web services for every activity in a business workflow, such as Web Services Business Process Execution Language (WS-BPEL) [1], designed to optimize the global QoS as a function of customer preferences and constraints. This is well known NP-hard optimization problem. Numerous researches have discussed requirements of QoS for endways systems addressing this issue [7, 8, 9]. The composition of service is the process by which new services get added to existing ones as per the functional requirement of the application. In this process for adding new services, various options may be available with functionally similar services, at that time non-functional parameters will be selection criteria.

The decentralized execution of composite service performs partitioning of programs scripted in BPEL into a number of subprograms that are smaller than the original script. Each subprogram executes on a distinct node or different BPEL node/server [10]. Fig. 1 illustrates a decentralized execution of a BPEL scripted programme. This is partitioned into five subprograms. Then subprograms are deployed on isolated nodes/servers, BS1 to BS5. The communication between them happens by asynchronous messaging. This decentralized composite service's execution is effective for improvement in QoS parameters like throughput and response time of composite service.

The work presented here partitioning a BPEL data-intensive process using a genetic programming approach [11]. The topology used for execution is distributed. It is based on numbers represented in fuzzy logic, where integrated processing and measurable service composition are performed using communication latency and costs within and between partitions. Optimization of service composition is performed by utilizing decision-making through multiple attribute groups and evaluation of cut-set matrix. In [12], using a combination of case analysis and simulations, verify the reliability of the algorithm. [13], this technique allows partitioning well-structured and unstructured processes by using graph transformations based on the representation of process structure graphs.

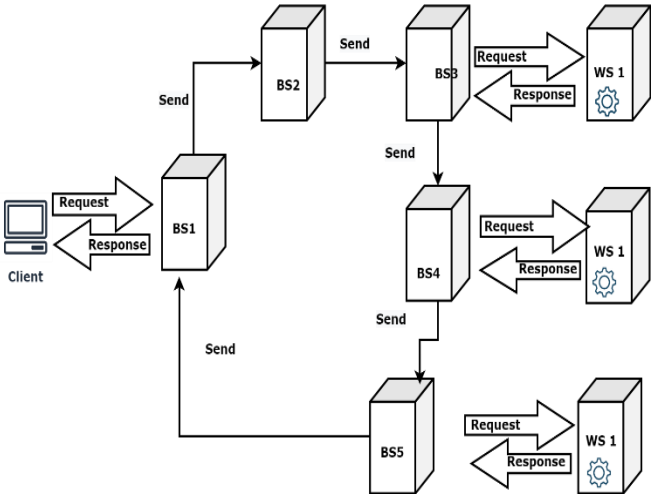


Fig. 1. Decentralized Topology for Execution of Composite Service in Composition.

Partitioning acts as the bridge between ideal and useful parallelism. Program dependencies reveal ideal parallelism through their control and data parallelism [14]. Any two-activity executions that are not related by control or data dependencies either directly or indirectly can be executed in parallel. The parallelism that can be used on a multiprocessor system is a subset of ideal parallelism [15]. Graph partitioning with optimization of performance is an NP-Hard problem in terms of computational complexity [16].

In the above-surveyed papers, no work is focused on assigning multiple high-cost nodes on a single server. The work done in [16] is based on real-time scenarios but any execution topology is unclear. Also, most works on decentralized execution [17] presented hypothetical data and hence the real-time execution of web services composition is needed to increase its acceptance in real-time use. The objectives of the paper are mentioned below.

1) In this paper, the BPEL program can be classified into High-Cost Task (HCT), and Low-Cost Task (LCT) statements.

2) In which, execution of more than one HCTs can be possible on a single server node and comprise the INVOKE, REPLY and RECEIVE statements.

3) The REPLY and RECEIVE statements are mostly kept at the last server node and the first node respectively from where client requests for composite service.

4) The INVOKE statements must be kept for execution on the corresponding web service or nearest location of web service as per availability of load capacity of the server node.

5) A LCT can be kept for execution on any server. LCT tasks are ASSIGN, SEND, IF, etc.

The following is how the rest of the paper is organized: The section two elaborates on related/previous work. Section III describes the proposed mathematically constructed algorithm flow. Section IV discusses the results analysis and discussion. Section V is where the paper's conclusion is found.

II. Related Work

This section paper reviews existing methods and frameworks that researchers have used to partition and execute Composite Services of Composition using a Decentralized Environment.

The author in [10], first introduced the BPEL program execution in the decentralized environment as Program Partitioning Problem. A program dependence graph is used with a set of portable and fixed types of activities. MDU, as well as the PGM heuristic approach, are proposed for partitioning the program [18]. The optimal partitions are determined using the MDU algorithm. The best partition is a partition that executes portable tasks to optimize (maximize) throughput. PGM algorithm is used to reduce the data on the server and the total number of messages to overcome the large computation time of MDU.

As [19], anticipated, Domain-Specific selection of service operates at the communication level which facilitates the application. However, there is a nonexistence of interoperability. Availability, accuracy, response time, and throughput, are considered QoS non-functional parameters. [20] the proposed algorithm is evolutionary which is employed to discover the best-fitting composition of service if the number of services is one. However, recursive processes are implemented when the number of services is one. Different selection strategies will choose different tasks.

Author in [21], survey the techniques of slicing programs that are recycled for performing according to the Program Dependency Graph (PDG).PDG is the greatest common approach in this survey and is effective since it handles data and control dependencies. In BPEL program tasks need allocation of specific servers for execution, due to this existing PDC-based approaches imperfect to apply to the problem of the BPEL program partition.

Author in [22], presents a method for realizing data flows in decentralized Internet of Things (IoT) systems based upon DX-MAN semantics. The algebraic semantics of the model enables direct data links between service producers and consumers of data. As a result, the data space of a decentralized environment is used to write data in and read data out. The authors validated the approach by means of the Blockchain of smart contracts. Results indicate approach is scalable with the growth in IoT systems size.

In [23] traditional techniques for optimization like Multidimensional Multiple-Choice Knapsack(MMCK) and Integer Linear Programming (ILP) are presented to report the problem of QoS-WSC. However, due to time complexity being exponential, these approaches have limited scalability when the problem size is small. To overcome these problems, approximate algorithms based on evolutionary search are proposed to find an optimal solution. Evolutionary computation methods such: Genetic Algorithm (GA) [24]–[26], Ant Colony Optimization (ACO) [27, 28], and Particle Swarm Optimization (PSO) [29], with an extraordinary amount of web-based services and various QoS aspects, algorithms were recycled to catch the service composition with optimal solution plan within a practically undersized period of interval.

By combining all objectives with a function for fitness these methods decrease to a single-objective problem (such as weighted sum techniques and fraction-based techniques).

Author [30], presents the hyper-cube peer-to-peer topology based on distributed system architecture. Efforts are made toward improving the average time and throughput of BPEL processes through the use of decentralized algorithms. The presented algorithms are supported a given BPEL process with decomposition. The presented approach provides a monitoring mechanism. Based on the experimental results, they discovered that the proposed architecture is better suited for long-running, data-intensive processes.

Using typed digraphs and a graph transformation technique [17], propose a technique for creating decentralized service compositions. They discuss the topology and interaction characteristics of the solutions. Based on experimentation, the authors describe a method for ranking topologies. Decentralized compositions are said to have low response times and high throughput on average.

Author [31], provides decentralized execution environments that optimize BPEL-based business processes through the use of shared spaces that represent a communication network among agents (intelligent) and a set of agents (cooperative) to execution of shared services.

Integer linear programming (ILP) takes existed widely used towards address the composition of services issue [32], Pareto dominance [33],QoS constraints decomposition [34], [35], reinforcement learning (RL) [36], or a combination of various techniques. Many evolutionary computation-based algorithms have been developed in recent years [37], [38] and swarm intelligence(SI) based [39], [40]. Compositional approaches that are QoS aware obligate remained proposed so that service compositions for the near-to-optimal solution can be found fairly quickly. The services composition with swarm-based intelligence and evolutionary-based computation approaches are presented in [40], [41], [42]. Researchers propose to achieve a trade-off between service composition with a near-optimal solution and a condensed computation time while achieving compositions of service in standings of QoS parameter optimality.

Meta-heuristics are also called approximation algorithms since they seek to discover the search space using various methods [43]. They do not recompense unusual consideration of optimization problems' mathematical environment or special experience involved with them. In [44], used Genetic Algorithms (GAs) to resolve the web-based service composition problem and demonstrated that GAs can be recycled efficiently for optimizing the composition of web-based services. Furthermore, expanding multiple objectives GA, [45], obtained adequate solutions in an undersized time. However, the difficulty of worst-case GA remains exponential for time complexity, which cannot be used for large applications where scaling is high.

It is complex enough to choose services based on multiple criteria even when one focuses on web composite service with a single user taking place in the pipeline. Few researchers proposed techniques, [46], [47] pay focus on the multi-attribute combinatorial auction (iterative) between various providers for services, whereas [48] pursues to advance these techniques by means of motivation device. We examined the user's QoS preference in the Big Data space, along with their service trust, in order to select services. Recently, [49], analyzed the problem of service composition from the perspective of a general Pareto-optimality to shrink the service composition's search space. An approximation estimate of the Pareto principle of optimality in polynomial time remained used in [50].

Author [51], presents customer authentic cost calculation method named Integrated Multi-Level Composite Service Model. They demonstrated the application of customer management in composition. However, lack in addressing quality parameters. In [52], researchers introduce blockchain architecture for the semantic composition of web services and develop QoS aware algorithm in terms of accuracy but considered QoS like throughput and response time.

The survey shows that most of the existing research works did not handle the optimality issue of service composition problem. Some of the research work on service composition has not even reached up to the phase of execution of composite service which is a major outcome of composition workflows. Therefore, an opportunity for research work on this breach is suitable.

III. Proposed Algorithm

The current section proposed a Multi High Tasks Genetic Algorithm (MHGA) technique which is novel and based on the allocation of multiple high-cost tasks on a single server node for the execution of composite service. This research discovers the usage of the improved genetic algorithm to address the problem of partitioning the BPEL program [13, 20, 22, 24, 26, 37, 44]. In addition, at hand is the absence of an evaluation or assessment model for simulation techniques of the BPEL program for the partitioning problem. This work develops a novel simulation model which is evaluating the effectiveness and efficiency of the new Multi High Tasks Genetic Algorithm. Here presents the proposed MHGA approach architecture and layout.

The BPEL development is self-possessed of a set of declarations called activities. These activities can remain categorized as high-cost activities and low-cost activities. High-cost activities, such as RECEIVE, REPLY, and INVOKE can be deployed and executed particular or nearest server node or engine of workflow. Any workflow engine can be assigned low-cost activities like SEND, ASSIGN, and IF.

A. The Description of Problem

The description of the BPEL program partitioning problem: $G = (HCT, LCT, DD, CD)$, is the program dependency graph is input as shown in Fig. 2. Where, $= \{h_1, h_2, .. h_n\}$, n stands a number of High-cost jobs in program. $LCT = \{l_1, l_2, .. l_m\}$, m stands a number of Low-cost jobs in the program.

A set of data dependencies. $DD = \{< v_i, v_j > \| v_i, v_j \in HCT \cup LCT\}$. A set of control dependencies $CD =$

$\{< v_i, v_j, Cd_p > \parallel v_i, v_j \in HCT \cup LCT \text{ and } Cd_p \in B\{boolean\ values\}\}$.

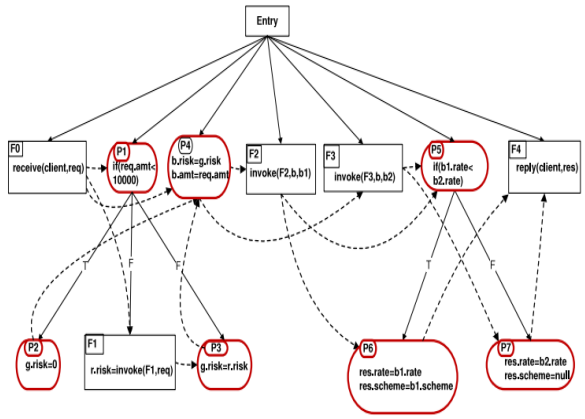


Fig. 2. BPEL Program Partitioning with Dependency Graph Example.

The computation cost on each n server (S) for tasks containing $\{receive, reply, invoke, send, assign\ if\}$ shown by $\{Cost_{receive}(n), Cost_{reply}(n), Cost_{invoke}(n), Cost_{send}(n), Cost_{assign}(n), Cost_{if}(n)\}$.

Throughput of the server is calculated by the server capacity and cost.

$$TH(S_n) = \frac{Capacity_{(n)}}{Cost_{(n)}} \quad (1)$$

Where $TH(S_n)$ is the throughput of the server. $Cost(n)$ is the total computational costs of tasks allotted on server S_n.

$$Cost(n) = \begin{Bmatrix} RE(n) * Cost_{receiver}(n) + RP(n) * Cost_{reply}(n) \\ +IV(n) * Cost_{invoke}(n) + SD(n) * Cost_{send}(n) \\ +AS(n) * Cost_{assign}(n) + IF(n) * Cost_{if}(n) \end{Bmatrix} \quad (2)$$

$Cost(n)$ is total cost on server S_n, where RE, RP, IV, SD, AS, IF are number of receive, reply, invoke, assign, send, if respectively.

$$F_{obj}(X) = min(TH(S_n)) \quad (3)$$

Throughput of plan X, which is minimum throughput among all servers in plan. The output of the plan X such that, $X = \{< L_m, H_n > \parallel H_n \in HCT, L_m \in LCT\}$. LCT and HCT assigned on partition such that Fitness is Maximal i.e., $F_{obj}(X)$ as well as precedence and control dependency constraints are satisfied.

B. Proposed MHGA Technique

The proposed technique is based on allocation of multiple (i.e. more than one) numbers of high cost tasks on single server nodes unlike any previous approaches. While allocation of high cost tasks on server nodes, available capacity of node is considered and checked for eligibility of placement of tasks on node using equation (4), if eq. (4) is true then only allocation of tasks is done on specific nodes. Here required capacity is nothing but cost of high cost task labeled as HCT with as current high cost task Otherwise node is allowed for execution of single task.

$$Available(Capacity(S_n)) \geq Required(Capacity(H_n)) \quad (4)$$

The model shown Fig. 3 is used to represent placement of lost cost and high cost task in solution. Where the integer number below LCT represents the number of low cost tasks. The number below HCT in representation is used for showing the number of high cost tasks allowed for execution on a particular server node.

C. Algorithm

The algorithm is proposed as in Algorithm 1. In step 1 algorithm generates solution with keeping LCT and HCT on server with respective partition. Next step calculates fitness using F_obj (X). Any two best solutions get selected for crossover process. In crossover bits from solution get cross with another solution in selected pair. A mutation operation get performs on offspring generated from crossover. These newly generated solutions get updated in original population. Fitness for updated solution population gets calculated for the selection of population for crossover and mutation in next iteration. This process will get repeated over number of iterations.

Algorithm 1: Program partitioning-genetic algorithm

Input: Program Dependency Graph, Data Dependency Set and Control Dependency Set

Output: Partition Plan X

Initialization: Initialize all population with random solution plan.
Each LCT is placed on any partition.
Each HCT is placed on partition so that maximum capacity will not exceed up to maximum two HCT on each server.
$n \leftarrow 0$;
While $n \leq populationsize$ ***do***
    Calculate fitness of each solution using following fitness formula by Eqn. (3)
    Calculate throughput of the server S_n by Eqn. (1)
    Calculate the cost of the server S_n by Eqn. (2)
    $n = n + 1$;
    end
$Max_{interation}$ =200;
while $Max_{interation}$ ***do***
    Selection: best two population with highest fitness will be selected for next step;
    Crossover: operation with 0.9 probability iscross from two population;
    new two populations get updated from initial population ;
    Mutation: operation with 0.1 probabilities is mutated. one bit from each new solution population gets changed to other value;
    Calculate fitness for updated solution population;
    end
return best execution plan devising the greatest value of fitness;

D. Model of Solution Plan

As shown in Fig. 3, a model of the population as solution plan is represented as Array List with a total of a number of LCTs (m) and a number of HCT's(n).The value in Array List indicates allotted LCT for first m elements and then next n elements allotted HCT in solution as population. An m stands for the total quantity of LCT, n stands for the total quantity of HCTs in solution plan. The value of s ranges from 0 to a total number of servers minus 1 (i.e. s_n-1, wherever s_n stands for the total quantity of servers) for allotment.

Population as solution plan									
LCT_1	LCT_2	LCT_3	LCT_4	...	LCT_m	HCT_1	HCT_2	...	HCT_n
s	s	s	s	...	s	s	s	...	s

m : total number of LCT

n : total number of HCT

s : range from [0 to total number of servers -1] for allotment

Fig. 3. Model of Population as Solution Plan.

IV. Results and Discussion

The proposed work is simulated on the Window 7 64 bits operating system. To carry out this work at least 4GB Ram and Intel Core i3processors are required. The Java Platform (JDK1.8) is used to design a model and calculate the performance parameters. The work focused on analyzing the simulation of proposed techniques (MHGA) for the partition of the BPEL program generated from composite service in service composition.

The proposed MHGA technique performance is compared with the three existing heuristic GA [13], [22], MDU, and PGM [10] techniques. In our simulation, the range of low-cost tasks (LCT) and high-cost tasks (HCT) are varying from 07:05 to 70:50. During the evaluation, computation time is observed in milliseconds keeping the population size of particles fixed to 20and the number of iterations to 200. Other GA parameters setting are crossover =0.9(probability), mutation =0.1 (probability) so that probability of crossover + mutation =1.0. The cost value for receive, reply, invoke, send, assign, if are taken as 0.6, 0.45, 2.5, 0.5, 0.6, 0.6 respectively from benchmark [10].

Fig. 4 shows the performance of the proposed (Multi High Tasks Genetic Algorithm) MHGA algorithm and GA algorithm from existing work. From Fig. 4 and Table I it is observed that the proposed algorithm partitioned the BPEL program for complex applications with high tasks in reasonable time similar to the GA algorithm for varying numbers of LCTs and HCTs. But through a new approach resource (server) utilization will be better and topology with a smaller quantity of server node scan be simple than existing approaches. As existing approaches use a total number of server nodes in topology equals to the quantity of HCT in the program which is very tedious and impractical for implementation especially in the case of the high number of HCTs.

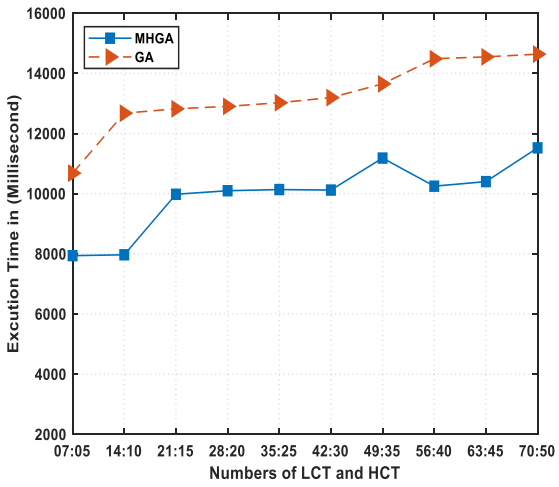


Fig. 4. Comparison of Execution Time between Proposed MGHA Approach with Existing GA with respect to Number of LCT and HCT.

TABLE I. Comparison of Time for Proposed MGHA Approach with Existing GA

Sr. No.	No. of LCT	No. of HCT	Time for MHGA	Time for GA
1	7	5	7940	10682
2	14	10	7968	12675
3	21	15	9980	12821
4	28	20	10098	12902
5	35	25	10137	13024
6	42	30	10120	13194
7	49	35	11184	13650
8	56	40	10249	14485
9	63	45	10401	14548
10	70	50	11523	14639

Our approach reduces the number of server nodes to up to half of the HCTs in the program which is practically possible for implementation. Fig. 5 shows the number of server nodes allotted for the proposed MHGA algorithm vs GA algorithm from existing work. From Fig. 5, it is observed that the proposed algorithm partitions the BPEL program within a smaller number of server nodes than the existing approach.

The simulation results are brief in Fig. 6. As shown, the average amount of throughput (request/seconds) of the business workflow beneath the partitioning solution plan found by the proposed MHGA and GA technique. The proposed MHGA techniques show a higher throughput as compared to the GA techniques.

It can be observed from Fig. 7, the computation time comparison among the proposed MHGA, GA, MDU, and PGM. The proposed MHGA techniques required less computation time as compared to existing techniques such as GA, MDU, and PGM. The proposed techniques performed better with the low and high-cost tasks. The PGM and GA technique also employs much fewer computation times than MDU in the attainment of a solution plan. GA technique also performed better than the existing MDU and PGM techniques. More specifically, MDU and PGM techniques are not addressing optimality issues of

optimization problems like GA-based technique hence detailed comparison for execution time with these approaches is not shown in this work.

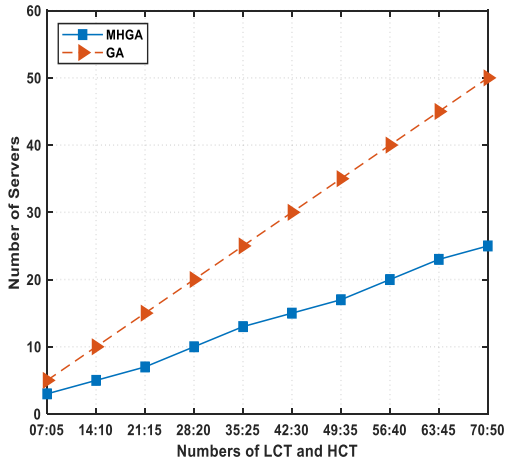


Fig. 5. Comparison of Number of Servers Utilized between Proposed MGHA approach with Existing GA with respect to Number of LCT and HCT.

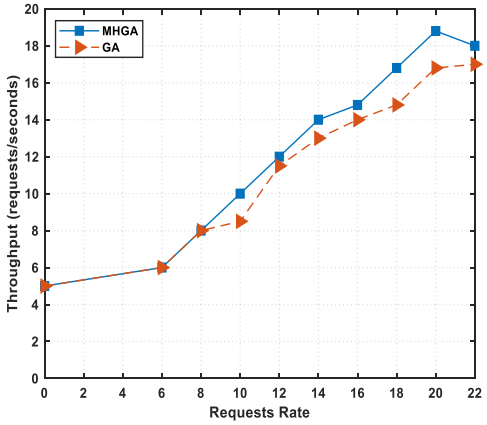


Fig. 6. Comparison of Throughput between Proposed MGHA approach with Existing GA with respect to Request Rate.

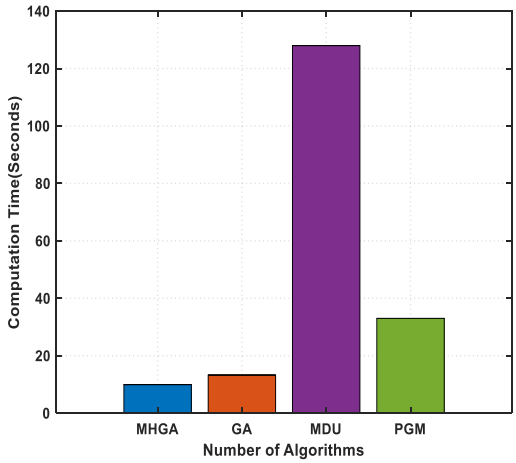


Fig. 7. Comparative Analysis of Average Computation Time between Proposed MGHA approach with Existing GA, MDU and PGM.

V. Conclusion

This paper performed a simulation of BPEL partition using a genetic algorithm through multiple high tasks allocation to a single server node. This work analyzes the existing genetic algorithms for the solution, execution time, and their nature as well as their number of server node requirements for actual execution. From the simulation results, it can be seen that the proposed MHGA technique is suitable for partitioning a large number of tasks in BPEL programs. The server node requirement in terms of quantity for the proposed approach is less than the existing ones; hence execution topology for large composite applications will be simple through the proposed approach. The proposed MHGA technique performed better than the existing GA, MDU, and PGM techniques in terms of the execution time, amount of server requests, and throughput.

In future work, consideration of issues like control dependencies and data dependencies can be explored in more detail for real-time data and complex application scenarios in dynamic environments.

References

[1] X. Chen, Z. Zheng, X. Liu, Z. Huang, and H. Sun, "Personalized QoS-Aware Web Service Recommendation and Visualization," IEEE Trans. Serv. Comput., vol. 6, no. 1, pp. 35–47, 2013, doi: 10.1109/TSC.2011.35.

[2] M. B. Juric, B. Mathew, and P. G. Sarang, Business process execution language for web services: an architect and developer's guide to orchestrating web services using BPEL4WS. Packt Publishing Ltd, 2006.

[3] D. Wutke, D. Martin, and F. Leymann, "A method for partitioning BPEL processes for decentralized execution.," in ZEUS, 2009, pp. 109–114.

[4] L.-J. Zhang, J. Zhang, and H. Cai, Services computing. Springer, 2007.

[5] L. Qi, Y. Tang, W. Dou, and J. Chen, "Combining Local Optimization and Enumeration for QoS-aware Web Service Composition," in 2010 IEEE International Conference on Web Services, Miami, FL, USA, Jul. 2010, pp. 34–41. doi: 10.1109/ICWS.2010.62.

[6] L. Barakat, S. Miles, I. Poernomo, and M. Luck, "Efficient Multi-granularity Service Composition," in 2011 IEEE International Conference on Web Services, Washington, DC, USA, Jul. 2011, pp. 227–234. doi: 10.1109/ICWS.2011.25.

[7] M. Alrifai and T. Risse, "Combining global optimization with local selection for efficient QoS-aware service composition," in Proceedings of the 18th international conference on World wide web - WWW '09, Madrid, Spain, 2009, p. 881. doi: 10.1145/1526709.1526828.

[8] W. Ahmed, Y. Wu, and W. Zheng, "Response Time Based Optimal Web Service Selection," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 2, pp. 551–561, Feb. 2015, doi: 10.1109/TPDS.2013.310.

[9] S.-Y. Hwang, C.-C. Hsu, and C.-H. Lee, "Service Selection for Web Services with Probabilistic QoS," IEEE Trans. Serv. Comput., vol. 8, no. 3, pp. 467–480, May 2015, doi: 10.1109/TSC.2014.2338851.

[10] M. G. Nanda, S. Chandra, and V. Sarkar, "Decentralizing execution of composite web services," in Proceedings of the 19th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, Vancouver BC Canada, Oct. 2004, pp. 170–187. doi: 10.1145/1028976.1028991.

[11] Z. Brahmi and I. Feddaoui, "Decentralized orchestration of BPEL processes based on shared space," in 2015 6th International Conference on Information Systems and Economic Intelligence (SIIE), 2015, pp. 60–65.

[12] S. Zheng, D. Feng, and J. Yu, "The algorithm of Web services composition of group decision-making based on fuzzy numbers," in Proceedings of the 3rd International Conference on Computer Science and Application Engineering, 2019, pp. 1–7.

[13] Y. Yu, H. Ma, and M. Zhang, "A genetic programming approach to distributed execution of data-intensive web service compositions," in Proceedings of the Australasian Computer Science Week Multiconference, 2016, pp. 1–9.

[14] J. Ferrante, K. J. Ottenstein, and J. D. Warren, "The program dependence graph and its use in optimization," ACM Trans. Program. Lang. Syst. TOPLAS, vol. 9, no. 3, pp. 319–349, 1987.

[15] V. Sarkar, "Automatic partitioning of a program dependence graph into parallel tasks," IBM J. Res. Dev., vol. 35, no. 5.6, pp. 779–804, 1991.

[16] G. Xue, J. Liu, L. Wu, and S. Yao, "A graph based technique of process partitioning," J. Web Eng., pp. 121–140, 2018.

[17] M. Pantazoglou, I. Pogkas, and A. Tsalgatidou, "Decentralized enactment of BPEL processes," IEEE Trans. Serv. Comput., vol. 7, no. 2, pp. 184–197, 2013.

[18] T. Mohsni and Z. Brahmi, "Partitioning BPEL program for decentralized execution based on Swarm Intelligence".

[19] O. Moser, F. Rosenberg, and S. Dustdar, "Domain-Specific Service Selection for Composite Services," IEEE Trans. Softw. Eng., vol. 38, no. 4, pp. 828–843, Jul. 2012, doi: 10.1109/TSE.2011.43.

[20] S.-C. Liu and S.-S. Weng, "Applying genetic algorithm to select web services based on workflow quality of service," J. Electron. Commer. Res., vol. 13, no. 2, p. 157, 2012.

[21] Y. Katsuno and H. Takahashi, "An Automated Parallel Approach for Rapid Deployment of Composite Application Servers," in 2015 IEEE International Conference on Cloud Engineering, Tempe, AZ, USA, Mar. 2015, pp. 126–134. doi: 10.1109/IC2E.2015.16.

[22] L. Ai, M. Tang, and C. Fidge, "Partitioning composite web services for decentralized execution using a genetic algorithm," Future Gener. Comput. Syst., vol. 27, no. 2, pp. 157–172, 2011.

[23] Y. Shi and X. Chen, "A Survey on QoS-aware Web Service Composition," in 2011 Third International Conference on Multimedia Information Networking and Security, Shanghai, China, Nov. 2011, pp. 283–287. doi: 10.1109/MINES.2011.118.

[24] G. Canfora, M. Di Penta, R. Esposito, and M. L. Villani, "An approach for QoS-aware service composition based on genetic algorithms," in Proceedings of the 2005 conference on Genetic and evolutionary computation - GECCO '05, Washington DC, USA, 2005, p. 1069. doi: 10.1145/1068009.1068189.

[25] W.-C. Chang, C.-S. Wu, and C. Chang, "Optimizing dynamic web service component composition by using evolutionary algorithms," in The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005, pp. 708–711.

[26] M. C. Jaeger and G. Mühl, "QoS-based selection of services: The implementation of a genetic algorithm," in Communication in Distributed Systems-15. ITG/GI Symposium, 2007, pp. 1–12.

[27] L. Aziz, S. Raghay, H. Aznaoui, and A. Jamali, "A new approach based on a genetic algorithm and an agent cluster head to optimize energy in Wireless Sensor Networks," in 2016 international conference on information technology for organizations development (IT4OD), 2016, pp. 1–5.

[28] Q. Wu and Q. Zhu, "Transactional and QoS-aware dynamic service composition based on ant colony optimization," Future Gener. Comput. Syst., vol. 29, no. 5, pp. 1112–1119, Jul. 2013, doi: 10.1016/j.future.2012.12.010.

[29] W. Wang, Q. Sun, X. Zhao, and F. Yang, "An improved particle swarm optimization algorithm for QoS-aware web service selection in service oriented communication," Int. J. Comput. Intell. Syst., vol. 3, no. sup01, pp. 18–30, 2010.

[30] D. Arellanes and K.-K. Lau, "Decentralized Data Flows in Algebraic Service Compositions for the Scalability of IoT Systems," in 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), Limerick, Ireland, Apr. 2019, pp. 668–673. doi: 10.1109/WF-IoT.2019.8767238.

[31] G. Xue, D. Liu, J. Liu, and S. Yao, "A process partitioning technique for constructing decentralized web service compositions," Softw. Pract. Exp., vol. 49, no. 10, pp. 1550–1570, 2019.

[32] V. Gabrel, M. Manouvrier, and C. Murat, "Web services composition: Complexity and models," Discrete Appl. Math., vol. 196, pp. 100–114, Dec. 2015, doi: 10.1016/j.dam.2014.10.020.

[33] Q. Yu and A. Bouguettaya, "Efficient Service Skyline Computation for Composite Service Selection," IEEE Trans. Knowl. Data Eng., vol. 25, no. 4, pp. 776–789, Apr. 2013, doi: 10.1109/TKDE.2011.268.

[34] S. X. Sun and J. Zhao, "A decomposition-based approach for service composition with global QoS guarantees," Inf. Sci., vol. 199, pp. 138–153, Sep. 2012, doi: 10.1016/j.ins.2012.02.061.

[35] H. Wang, P. Ma, Q. Yu, D. Yang, J. Li, and H. Fei, "Combining quantitative constraints with qualitative preferences for effective non-functional properties-aware service composition," J. Parallel Distrib. Comput., vol. 100, pp. 71–84, Feb. 2017, doi: 10.1016/j.jpdc.2016.10.013.

[36] H. Wang, M. Gu, Q. Yu, H. Fei, J. Li, and Y. Tao, "Large-Scale and Adaptive Service Composition Using Deep Reinforcement Learning," in Service-Oriented Computing, vol. 10601, M. Maximilien, A. Vallecillo, J. Wang, and M. Oriol, Eds. Cham: Springer International Publishing, 2017, pp. 383–391. doi: 10.1007/978-3-319-69035-3_27.

[37] A. S. da Silva, H. Ma, and M. Zhang, "Genetic programming for QoS-aware web service composition and selection," Soft Comput., vol. 20, no. 10, pp. 3851–3867, Oct. 2016, doi: 10.1007/s00500-016-2096-z.

[38] F. Wagner, F. Ishikawa, and S. Honiden, "Robust Service Compositions with Functional and Location Diversity," IEEE Trans. Serv. Comput., vol. 9, no. 2, pp. 277–290, Mar. 2016, doi: 10.1109/TSC.2013.2295791.

[39] M. S. Hossain, M. Moniruzzaman, G. Muhammad, A. Ghoneim, and A. Alamri, "Big Data-Driven Service Composition Using Parallel Clustered Particle Swarm Optimization in Mobile Environment," IEEE Trans. Serv. Comput., vol. 9, no. 5, pp. 806–817, Sep. 2016, doi: 10.1109/TSC.2016.2598335.

[40] X. Xu, Z. Liu, Z. Wang, Q. Z. Sheng, J. Yu, and X. Wang, "S-ABC: A paradigm of service domain-oriented artificial bee colony algorithms for service selection and composition," Future Gener. Comput. Syst., vol. 68, pp. 304–319, Mar. 2017, doi: 10.1016/j.future.2016.09.008.

[41] C. Jatoth, G. R. Gangadharan, and R. Buyya, "Computational Intelligence Based QoS-Aware Web Service Composition: A Systematic Literature Review," IEEE Trans. Serv. Comput., vol. 10, no. 3, pp. 475–492, May 2017, doi: 10.1109/TSC.2015.2473840.

[42] S. Mistry, A. Bouguettaya, H. Dong, and A. K. Qin, "Metaheuristic Optimization for Long-term IaaS Service Composition," IEEE Trans. Serv. Comput., vol. 11, no. 1, pp. 131–143, Jan. 2018, doi: 10.1109/TSC.2016.2542068.

[43] M. Gendreau and J.-Y. Potvin, "Metaheuristics in Combinatorial Optimization," Ann. Oper. Res., vol. 140, no. 1, pp. 189–213, Nov. 2005, doi: 10.1007/s10479-005-3971-7.

[44] M. A. Amiri and H. Serajzadeh, "QoS aware web service composition based on genetic algorithm," in 2010 5th International Symposium on Telecommunications, 2010, pp. 502–507.

[45] H. Wada, J. Suzuki, Y. Yamano, and K. Oba, "E3: A Multiobjective Optimization Framework for SLA-Aware Service Composition," IEEE Trans. Serv. Comput., vol. 5, no. 3, pp. 358–372, 2012, doi: 10.1109/TSC.2011.6.

[46] Q. He, J. Yan, H. Jin, and Y. Yang, "Quality-Aware Service Selection for Service-Based Systems Based on Iterative Multi-Attribute Combinatorial Auction," IEEE Trans. Softw. Eng., vol. 40, no. 2, pp. 192–215, Feb. 2014, doi: 10.1109/TSE.2013.2297911.

[47] Y. Zhong, X. Li, and Q. He, "Iterative auction based service selection for multi-tenant service-based systems," in Proceedings of the Australasian Computer Science Week Multiconf a single web service may not have been sufficient for most complex business operations, erence, 2017, pp. 1–4.

[48] P. Wang and X. Du, "QoS-Aware Service Selection Using an Incentive Mechanism," IEEE Trans. Serv. Comput., vol. 12, no. 2, pp. 262–275, Mar. 2019, doi: 10.1109/TSC.2016.2602203.

[49] Y. Chen, J. Huang, C. Lin, and J. Hu, "A Partial Selection Methodology for Efficient QoS-Aware Service Composition," IEEE Trans. Serv. Comput., vol. 8, no. 3, pp. 384–397, May 2015, doi: 10.1109/TSC.2014.2381493.

[50] I. Trummer, B. Faltings, and W. Binder, "Multi-Objective Quality-Driven Service Selection—A Fully Polynomial Time Approximation Scheme," IEEE Trans. Softw. Eng., vol. 40, no. 2, pp. 167–191, Feb. 2014, doi: 10.1109/TSE.2013.61.

[51] K. Sudhakar, M. James Stephen, Trummer, B. Faltings, an "Cloud Oriented Integrated Composite Services over SOA in Distributed Computing," International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-10 Issue-3, February 202, pp. 1-7.

[52] S. Sridevi S, G. Karpagam, B. Vinoth, and J. Uma, "Investigation on Blockchain Technology for Web Service Composition: A Case Study" International Journal of Web Services Research Vol. 18, Issue no 1, 2021, pp. 70-93.

Design Processes for User Engagement with Mobile Health: A Systematic Review

Tochukwu Ikwunne, Lucy Hederman, P. J. Wall
ADAPT Centre, Trinity College Dublin, Ireland

***Abstract*—Despite the importance of user engagement in mHealth system efficacy, many such interventions fail to engage their users effectively. This paper provides a systematic review of 10 years of research (32 articles) on mHealth design interventions conducted between 2011 and 2020. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) model was used for this review with the IEEE, Medline EBSCO Host, ACM, and Springer databases searched for English language papers with the published range. The goal of this review was to find out which design process improves user engagement with mHealth in order to guide the development of future mHealth interventions. We discovered that the following six analytical themes influence user engagement: design goal, design target population, design method, design approach, socio-technical aspects, and design evaluation. These six analytical themes, as well as 16 other specific implementations derived from the reviewed articles, were included in a checklist designed to make designing, developing, and implementing mHealth systems easier. This study closes a gap in the literature by identifying a lack of consideration of socio-cultural contexts in the design of mHealth interventions and recommends that such socio-cultural contexts be considered and addressed in a systematic manner by identifying a design process for engaging users in mHealth interventions. Based on this, our systematic literature review recommends that a framework that captures the socio-cultural context of any mHealth implementation be refined or developed to support user engagement for mHealth.**

Keywords*—Design process; mobile health; socio-cultural; user-centered design; user engagement

I. Introduction

Many mHealth interventions fail to achieve or sustain their stated goals [1][2]. Many reasons have been put forward to explain this, one of which asserts that the effectiveness of mHealth initiatives is overly dependent on user engagement [3][4][5]. In this context, the concept of user engagement is critical. Although the term "user engagement" has several different interpretations, it is critical to have a common meaning of the term's specific definition because the various interpretations have led to a great deal of misunderstanding [6]. For the purposes of this study, user engagement should be defined as "the emotional, cognitive, and behavioral connection that exists between a user and a resource at any point in time and possibly over time" [7]. User engagement is critical in mHealth, with many researchers (e.g., [8][9]) arguing that the mHealth design process should take into account the needs of various users. However, many current mHealth interventions are based on pre-existing healthcare system constructs [10], encouraging designers to base their designs on assumptions that have not been validated with primary user input [11]. As a result, the interventions that result are less effective than those that include end-user needs [10] and input from relevant stakeholders such as commercial app industries and design experts [9]. The author in [11] defines user-centered design as a method that is informed by the needs and understanding of a particular end-user group and plays an important role in achieving user engagement with technology. People must engage in mHealth interventions for them to be effective, but engagement is frequently inadequate [12].

It has been stated that the effectiveness of mHealth initiatives is highly depended on user engagement [3]. However, despite the claimed importance of user engagement in mHealth system efficacy, many such interventions frequently lack user-engaging attributes [13]. According to [13], some mHealth apps lacked engaging and customizable features because the apps did not include any specific strategies to facilitate user engagement. Furthermore, [14] indicates that user engagement is a critical factor in determining the success of any mobile application.

Thus, additional research is warranted to improve understanding of user engaging features in technology in general, and mHealth, as well as to develop techniques and methodologies to facilitate and sustain user engagement [15]. As discussed in [11], it is also important to consider the socio-cultural contexts associated with user engagement when attempting to achieve user engagement with technology. According to [16], a lack of engagement with mHealth systems is caused by socio-cultural and organizational issues, such as when mHealth applications developed in the Global North are implemented in the Global South, where there may be numerous social, cultural, and belief differences. In such cases, it is critical that implementation take users' socio-cultural contexts into account in order to improve mHealth systems. This point is emphasized further by [17], who claims that the assumption that technology developed in the Global North can simply be dropped into the Global South and expected to work is a "fallacy."

This systematic review identifies a gap in the literature by highlighting the lack of consideration of users' socio-cultural contexts in the design of mHealth interventions and proposes that such user group socio-cultural contexts be considered. This is because techno-centric approaches to mHealth design and user engagement that are solely focused on technology, as well as other approaches that rely on existing universal frameworks for user-centered design, have been shown to be ineffective [18].

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. grant 18/CRT/6222 at the ADAPT SFI Research Centre at Trinity College Dublin.

In this systematic review, we seek to present the process of designing engaging mHealth interventions, situating the design process within the context of a user-centered design framework, and contextualizing the results by also incorporating the design processes of other mHealth interventions. As their guiding approaches, these are explicitly named design approaches user-centered, human-centered design, double diamond, and Hasso-Plattner Institute. As a result, the paper emphasizes the importance of improving understanding of user-engaging features in technology in broad sense, and mHealth precisely, as well as the need of developing more robust techniques and methodologies to facilitate and sustain user engagement.

II. Literature Review

This review of the literature begins with a discussion of the body of work dealing with the success and failure of mHealth. The section then examines the body of work on designing mHealth for engaging experiences before emphasizing the importance of considering the fit between various conceptions of engagement and the design process in mHealth design.

A. Design Success and Failure in mHealth

In the field of mHealth, defining success or failure is a difficult task. What is clear is that implementing such systems in a sustainable and scalable manner is difficult. The author in [1] states that "most information systems in developing countries fail either completely or partially," while [19] states that "successful examples of computerisation can be found... but frustrating stories of systems that failed to fulfill their initial promise are more common" (p.1). The success of any new technology is dependent on its successful integration, diffusion, and long-term use by intended users, according to [20]. Projects have been shown to be productive when they are tailored to the local context and language [21], and when they are developed and implemented with the participation of local private service providers [21][22]. A number of studies have demonstrated that mHealth interventions in the Global South are useful, particularly in improving treatment adherence, appointment compliance, data collection, and the development of support networks for health workers [23][24][25]. Although it is acknowledged that mHealth, particularly in the Global South, has great potential, many mHealth systems have historically failed to deliver on their initial promises [19]. There are numerous challenges and risks associated with the design, implementation, and adoption of such systems [26][27].

There have been numerous explanations advanced for this high level of mHealth failure. One of the primary causes of the failure was identified in [28]. Socio-cultural considerations, according to the article, have a significant impact on the implementation of any health information system. They discussed about the socio-cultural issues that arise when health information systems are transferred between two African developing countries' public health sectors (Mozambique and South Africa). The article demonstrates that transferring between two countries involves issues such as cultural differences, adjustment, and adaptation. While the transfer was deemed successful, the health information system needed to be flexible enough to support local variations. Similarly, a number of articles [29][30][31][32] have argued that successful implementations necessitate a better understanding of user groups' sociocultural contexts due to their importance and impact on the scale and sustainability of mHealth initiatives.

A major cause of failure, according to [33], is an unsuitable design in relation to the needs and context of use. Similarly, taking a techno-centric approach to mHealth implementation without considering socio-technical issues, according to [34], can be detrimental. Another reason for mHealth failure is the use of a top-down approach by implementers [35]. This approach is techno-centric, with users having no control over the technology that they expect to use. Furthermore, many mHealth systems are designed, developed, and imported from the Global North. According to [17], assuming that such systems will fit into any Global South country without considering users is a "fallacy." This highlights the significance of creating engaging experiences for users of mHealth interventions.

B. Designing for Engaging mHealth Experience

Although designing for engaging experiences is a widely stated goal of interactive system development across many disciplines, there are no guidelines in place to communicate designers' efforts to make things engaging [36]. The problem has been exacerbated by the lack of a unified definition of engagement. It is difficult to know whether the systems we design are engaging or to identify which aspects of technology interaction engage or fail to engage users if user engagement is not understood [37]. There are several definitions of user engagement, and the various viewpoints have resulted in a great deal of misunderstanding [6]. It's unclear how valuable these viewpoints are to designers. [38]. The author in [6] defined user engagement as "the total set of user relationships toward IS and their development, implementation, and use" (p. 514). The psychological state of mind required by the user to enjoy the representation, i.e., a willing suspension of disbelief, has also been defined as "user engagement" [39]. Other points of view on user engagement have shifted the focus away from the individual user and toward the designer. The author in [40] investigated methods for attracting people and encouraging interaction. The author in [41] was interested in motivating and improving the user experience of the application, whereas [42] defined engagement relying on their synthesizing of esthetical, flows, enjoy, and information interaction theories, as well as previous work in the application areas of video games, web searching, and educational software. Other definitions include [43] user experience, spatiotemporal, compositional, and sensual "threads of experience."

An important but frequently overlooked aspect of engagement research is the fit between various conceptions of engagement and the design process. The evaluation of user engagement is critical in the design of engaging experiences. However, there is very little attention given to incorporating engagement measures into the design process [38], implying that an improved mHealth design process is required to strengthen user engagement [44].

III. Material and Methods

The literature review provides a comprehensive description of the current state of the mHealth design process from the standpoint of user engagement. As a result, it can be used to inform future research and studies in the mHealth design process as a means of improving user engagement with mHealth technologies.

A. Introduction

A systematic qualitative analysis was used to categorize data based on different analytical themes. Articles from 2011 to 2020 were searched in four electronic databases (IEEE, Medline EBSCO Host, ACM, and Springer Link). The time period 2011 to 2020 was chosen to ensure that relevant articles associated with existing design process frameworks to improve user engagement with mHealth technology were found. The search string and their combinations that were used include "design process OR design process framework", “design process evaluations AND mobile health", “mobile health OR mHealth”, “user engagement AND mHealth design process”, “user engagement AND health”, “mHealth AND interventions”, “mHealth AND design process”, “mHealth applications AND user engagement”. This ensures that the mHealth design process is covered broadly across disciplines such as health informatics, information technology, and human-computer interaction (HCI). The search and selection process are depicted in Fig. 1. According to Fig. 1, the search yielded 3700 articles, 3100 of which remained after duplicates were removed. Another 2496 articles were eliminated because they were either (1) not published in English, (2) lacked full text, or (3) did not discuss the design process of mHealth interventions. This step resulted in 604 distinct articles. A further 572 articles were eliminated based on the following criteria: types of interventions studied - articles that do not deal with health interventions are excluded; if they did not report measured outcomes such as performance-based measures, self-report measures, or clinician-reported measures; and if they were not peer-reviewed.

This leaves 32 articles for review. The majority of the 32 articles included in the review came from Medline, with 20 articles, Springer and ACM each having 5 articles, and IEEE having the fewest, with only two. These articles were examined using the set of analytical themes described in the following section.

B. Thematic Analysis and Coding Scheme

The analysis of 32 selected articles is guided by [45] “six model of thematic analysis.” Table I shows the themes used to categorize reviewed articles. According to [45], thematic analysis "provides a flexible and useful tool that can potentially provide a rich and detailed, yet complex account of data."

The following themes were generated in total: design process goal, design approach, whether socio-technical aspects of intervention were addressed, design methods, design target audience, scalability, and design validation and evaluation.

The six themes are generated in accordance with the [45] model.

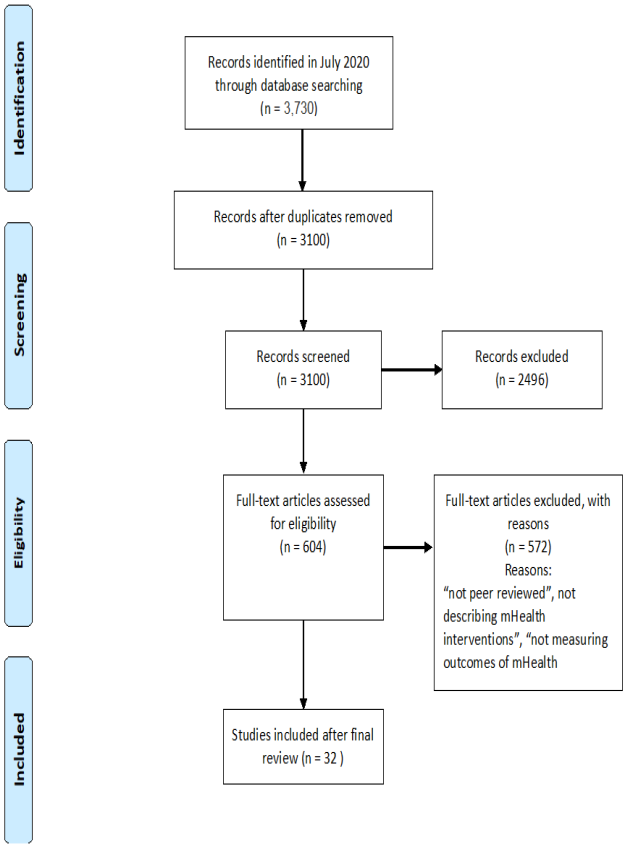


Fig. 1. The Article Selection Workflow.

In order to improve user engagement with mHealth interventions, the first step was to conduct a thorough review of the selected articles, address and analyze the articles, and keep in mind how the selected articles described the mHealth design process. The second step was to generate preliminary codes by highlighting phrases or sentences in the selected articles and creating shorthand labels (codes) to describe their content. Coding is the process of breaking down large amounts of data into smaller chunks of meaning. The codes were created and modified during the coding process without the use of pre-set codes. The coding was completed on an Excel sheet with the intention of not coding the entire data set because the primary concern is to address and analyze the data associating design process of improving user engagement. Two people coded the articles independently. Before moving on to the rest of the articles, each of the codes was compared, discussed, and modified in the third step. Despite the fact that not all of the text was coded, every coded reviewed article was relevant to or specifically addressed user engagement improvement. The fourth step was to look through the codes for a theme, idea, or concept that captured and summarized the data's meaningful and recurring patterns. According to [45], there are no strict guidelines about what constitutes a theme. A theme's significance defines it. The codes were scrutinized at this point to ensure that they fit together into a larger theme that addresses the design process of improving user engagement. The fifth, sixth, and final steps involved naming, reviewing, and refining codes to create the six themes.

TABLE I. THEMES USED TO CATEGORIZE REVIEWED ARTICLES

Themes	Definitions	Supporting Review Articles
Goal of the design	The theme design goals are the purposes for design work that are typically agreed upon by designers of such design work.	"The primary analysis is concerned with the marginal effect that is the average over time, of the contrast between the two possible intervention options. In secondary analysis, moderation with the goal of understanding in which circumstances one intervention option is more effect, can be explored". (Review 31)
Design target populations	Design target populations are users who mHealth designers consider when developing mHealth interventions.	"The study showed, however, that **patients** were satisfied with the phone application and it improved on their self-reported depressive symptoms" (Review 6)
Design methods	Design methods are techniques or tools for design work that provide a variety of activities that a designer may use as part of the overall design process.	We identified variables that corresponded to patient and scientific research priorities, discussed potential measurement schemes, and began to investigate technological options (eg, data streams, sensors, active tasks, analytical methods). We also started talking about a variety of technical, user experience, regulatory, and other issues related to the research program. (Review 32)
Design approach	The design approach refers to the solution-based method used in developing mHealth interventions.	"Applying **human-centred methods** in the design of e-health solutions requires that designers must take particular considerations when patients and healthcare professionals are involved in the design process." (Review 7)
Socio-technical aspects	The socio-technical aspects of interventions are defined as case-specific interventions based on qualitative and empirical evidence [46].	Furthermore, the overall organizational socioeconomic context of the clinical system setup must be investigated. (Review 2)
Design evaluations	"Evaluation is used to refer to measures taken, and analysis performed to assess (i) the interaction of users or a health system with the digital health intervention strategy, or (ii) changes attributable to the digital health intervention." [47]	The application and evaluation of this framework is demonstrated through the use case of a mHealth app that was designed to read the results of the tuberculin skin test, which is used to detect latent tuberculosis infection (LTBI) and for which a prototype was available. (Review 30)

Data associated with each theme were read to see if the data truly supported the theme and how the themes work within a single article as well as across all articles. The process of naming themes entailed giving each theme a short and simple name. As a result, we extracted themes from the reviewed articles until we determined that no more themes might be derived from the data. The following section delves into the specific meanings of the set of analytical themes.

C. Themes of Design Process for Improving user Engagement

Any design process, according to [48], is "the specific series of events, actions, or methods by which a procedure or set of procedures is followed, in order to achieve an intended purpose, goal, or outcome" (p. 408). Fig. 2 depicts the design process for increasing user engagement with mHealth interventions, which consists of six analytically generated themes, each of which can be refined by a number of descriptive themes. Design goal, design target audience, design methods, design approach, socio-technical aspects, and design evaluations are the six analytical themes.

These analytical themes, as well as their descriptive themes, are discussed in greater detail in the results section that follows.

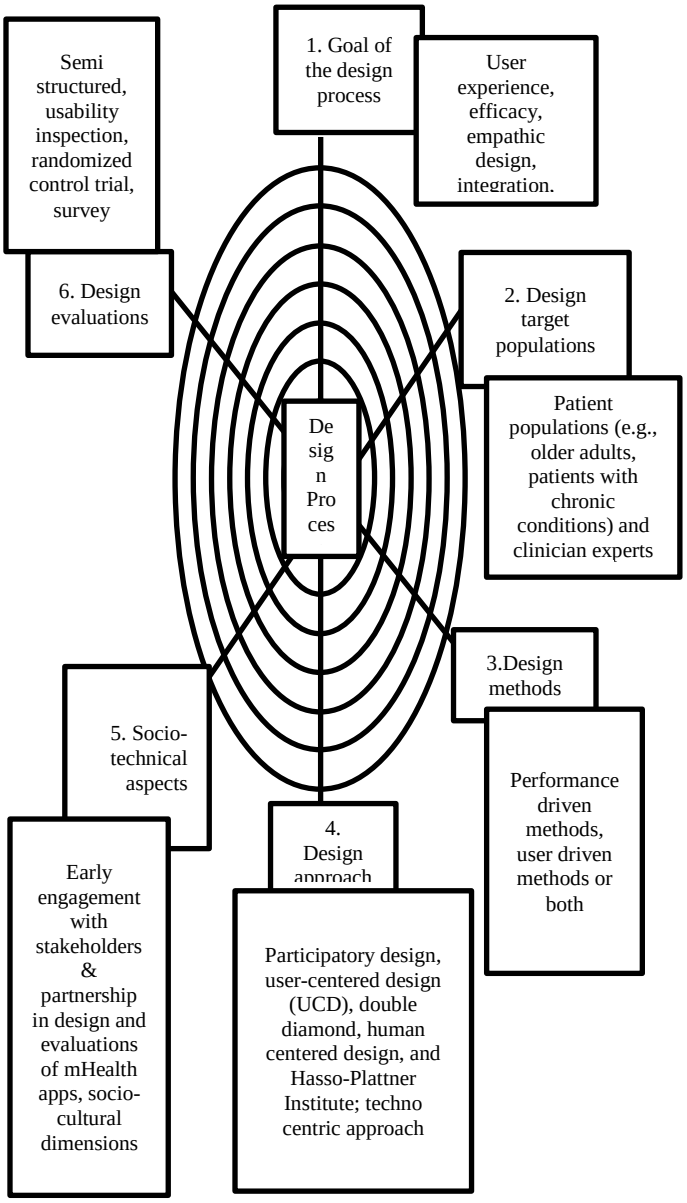


Fig. 2. Six Analytical Themes of Design Process.

IV. Results

The six analytical themes provide a thorough examination of the mHealth design process with the goal of increasing user engagement with mHealth technology. They reveal a wide range of trends and discoveries. To describe the design process that can improve user engagement with mHealth interventions, six analytical themes were developed, each of which can be explained by several descriptive themes.

A. Goal of the Design of mHealth Interventions

Although there are numerous publicly available mHealth interventions, particularly mental health apps, as identified in the reviewed articles, current knowledge about the goals for designing such interventions is limited, particularly from a sociotechnical and user-centered perspective [49]. As a result, six descriptive themes related to design goals were identified: user experience, efficacy, empathic design, integrated, predictive, personalized, and inclusive. These are now described in greater detail.

1) User expereience: The reviewed articles stressed that mHealth interventions should not be medically approved solely on the basis of their effectiveness; the experience of users during the use process should also be evaluated [49] [50]. User experience should be researched and incorporated into the intervention design process [51] [52][53]. Although having a poor usability mHealth intervention may not have a discernible negative impact on users, the negative experience of users may prevent them from accepting and approving new mHealth technological interventions in the future [54][55][56][57].

2) Efficacy: The efficacy of mHealth interventions was frequently lauded in the reviewed articles. Many of the mHealth interventions described in the reviewed articles had observable outcomes [58][59][60] [61][62][63][64]. Despite the fact that the efficacy of mHealth interventions varied significantly across studies, all of the reviews indicated that mHealth was a viable concept with the prospective to improve patient health.

3) Empathic design: According to the reviewed articles, it is critical to consider the user's feelings toward mHealth products; users quickly lose interest when their feelings about using products are jeopardized [65][66]. The preference was for mobile device users to be paid attention to their feelings toward mHealth products [67] and to receive feedback on continuous monitoring data on user emotions while using mHealth products, such as how their feelings progress over time with mHealth product use [53], predicted possible causes and solutions [68][69][70].

4) Integrated: An mHealth platform should not be regarded as a stand-alone tool to be used in isolation. Collaboration among practitioners, other healthcare service components, communities, caregivers, patients, and their dependents is required in various aspects of mHealth interventions [59][7].

5) Predictive and personalized: Users of mHealth interventions desired not only automatically tailored information, but also the ability to personalize the mHealth intervention. The reviewed articles emphasized the importance of mHealth intervention users being able to choose when and how they receive SMS messages [72], setting goals for future use of the mHealth tool to personalized lifestyle with synchronous communication with a health care professional [73], and participating in identifying the mHealth system requirements [74][75].

6) Inclusive: According to the reviewed articles, inclusive health care systems based on mobile interventions (for example, in mental health) have frequently been viewed positively [76] in developing countries, [70][77] due to the obvious considerably large penetration rates of mobile technologies and the effectiveness of human resources. Furthermore, because of their intimate and confidential nature, mobile solutions can be effective in a culture that stigmatizes mental health issues [78]. The design processes examined in this review had the following goals: high-quality user experience, efficacy, empathic design, integration, "predictive and personalized," and inclusive. Fig. 3 depicts the distribution of design process goals in the articles reviewed.

The analysis of the design goals revealed that all mHealth intervention designs, in one way or another, attempt to achieve some goals. It was discovered that none of the mHealth intervention designs addressed a combination of all the goals used in this study. The most frequently addressed design goal is efficacy (59%).

In 17 studies, user experience is the second most implemented design goal (53 percent). With 12 (38 %) and 11 (34%), respectively, studies, inclusiveness and empathic design process ranked third and fourth in terms of most used design goals. With a total of 7 (22 %) studies and 5 (16 %) studies, integration and predictive and personalized were ranked fifth and sixth, respectively.

B. Design Target Population

The users who mHealth designers consider when developing mHealth interventions are referred to as the design target population. The reviewed articles reported three design target populations: patient population (e.g., older adults, patients with chronic conditions) [69][71][79][80]; or both (patient and clinician experts) [52][53].

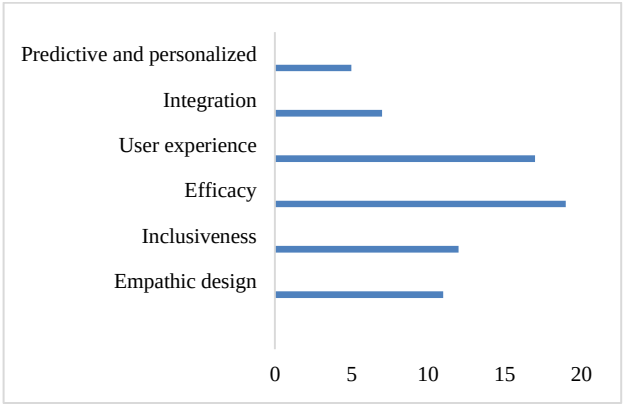


Fig. 3. mHealth Design Goals.

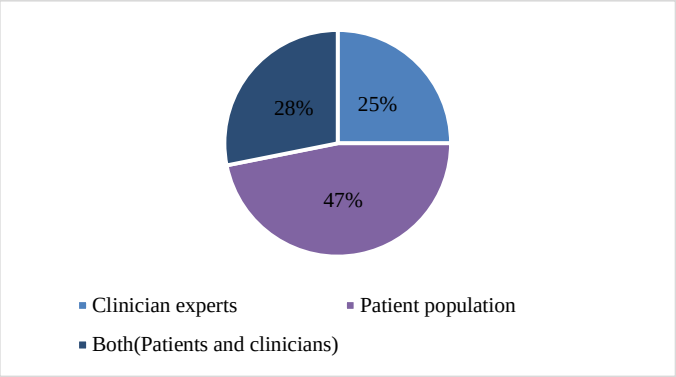


Fig. 4. mHealth Design Target Population.

The distribution of the articles across the mHealth design target population is depicted in Fig. 4. According to [81], mHealth projects frequently involve distinct stakeholders such as patient populations (e.g., older adults, patients with chronic conditions) and clinician experts, or both.

Eight studies (25%) targeted the population of clinician experts, while nine articles (28%) targeted the population of patients. Fifteen articles (47 %) were directed at both patients and clinicians.

C. Design Methods

Design methods are the methods or tools for designing that offer a variety of activities that a designer could use as part of an overall design process. Performance-driven methods, user-driven methods, and both were identified as three descriptive themes that could improve design methods (performance and user driven).

1) User driven methods: The reviewed articles underscored the importance of involving users from the start of the design process, to recognize, tap, and comprehend their explicit and implicit knowledge and ideas [50][65]. User-driven innovation methods range from casual observations to collaborations and intensive user participation in co-creation processes [71][79].

2) Performance driven methods: Performance driven methods entail the use of expert analyses to generate a continuous flow of improvement ideas that are strongly focused on the desired results. Performance-driven methods are commonly used when it is difficult to extract ideas or information from potential users. Among the performance methods extracted from the reviewed articles were using machine learning model for predicting patients' ambience, feelings, psychological states, actions, environmental factors, and social context; demonstrating that the mobile diary tool can increase client adherence to therapeutic activities; collecting users' psychological, physiological, and activity information for mental health research; and using assisted cognitive behavioral therapy for insomnia [59][60].

3) User driven and performance driven methods: Some of the reviewed mHealth interventions used both user-driven and performance-based design methods, such as [52] personal health monitoring and feedback system for bipolar disorder patients, [53] autonomous, intelligent mobility aid for older adults, [72] SMS-based application to motivate behavior change among tobacco users, and [82] remote measurement technologies (RMT) to study central nervous system function.

D. Comparative Analyses of the Design Methods, Design Goals and Design Target Population

In terms of the three descriptive themes of design methods used in the reviewed articles, user driven methods and both (user driven and performance driven methods) share common design goals, whereas performance driven methods do not (Fig. 5).

Fig. 6 depicts the comparative analyses of design methods and design target populations presented in the review. It should come as no surprise that there is a strong correlation between the targeted population and the design methods used in the mHealth intervention design process. In 7 (22 %) of the articles aimed at patients and clinicians, a combination of user-driven and performance-driven methods were used. The user-driven method was used in 17 (53%) of the reviewed articles that targeted both patients and clinicians, with 15 (47%) focusing on patient populations and 2 (6%) focusing on both patient and clinician populations. In 8 (2%) of the clinician-targeted articles, the performance-driven method was used.

User-driven methods (53%) were the most commonly used methods for the design process of mHealth interventions primarily aimed at patient populations, followed by performance-driven methods (25%) aimed at clinician experts, and finally a combination of user-driven and performance-driven methods (22%).

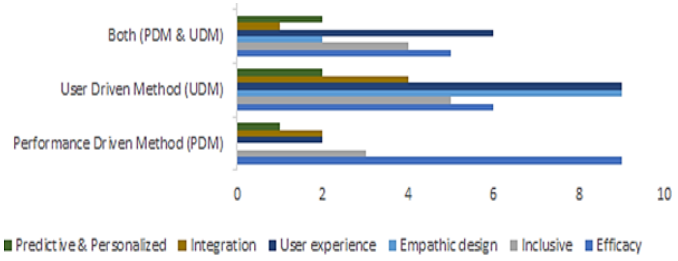


Fig. 5. Mobile Design Goals in different Design Methods.

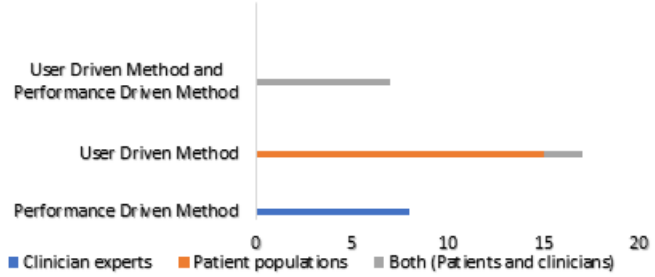


Fig. 6. Mobile Design Methods by Design Target Population.

E. Design Approach

The solution-based method used in designing mHealth interventions is referred to as the design approach. The following six approaches were used in the reviewed articles: participatory design, user-centered design (UCD), double diamond, human centered design, and Hasso-Plattner Institute; and techno centric approach.

1) Participatory Design (PD): Participatory design (PD) is an approach that involves all stakeholders in the intervention design process and ensures that the interventions developed are usable and meet the needs of users. The author in [83] defines participatory design experience as a transition in mindset and perception toward people from designing for participants to designing with participants. It is the notion anyone can contribute towards the design process in some way and that, given the right tools, they can be both articulate and creative. According to [84], timely engagement and partnership with stakeholders is critical for mHealth implementation. The concept of partnership implies shared goals, shared accountability for outcomes, distinct accountabilities, and reciprocal obligations. The PD process consists of nine steps: introduction, analyses, idea generation, idea selection, prototyping, testing, adjusting, implementation, and evaluation [85]. In the reviewed articles, the PD approach was used in identifying the system requirements for the design of interventions for schizophrenia [71], trait anxiety in adults [66], breast cancer [67], self-monitoring behavior [69][74] and Ebola preparedness [70].

2) User-Centered Design (UCD): UCD is a process that places users at the center of product design and development. Its primary goal is to make interactive products usable by analyzing system usage and applying human factors and usability information and methods [86]. UCD is comprised of four steps: (1) comprehend and specify the context of use; (2) specify the user and organizational specifications; (3) develop design solutions; and (4) evaluate design in relation to specifications [87]. The UCD approach was used in the reviewed articles to identify system requirements for the design of an intervention for multiple sclerosis [82], mental illness management [51][52], intelligent mobility aid for older adults [53], support for adolescents coping with chronic pain [61], behaviour patterns [80][88] and self-management [82][89].

3) Double Diamond (DD): The Design Council (2007) developed the Double Diamond approach, which is based on the application of design thinking in businesses and innovation designs and has four phases (discovery, definition, development, and delivery)[1]. There was limited use of the Double Diamond approach in the review articles, with only [77] using it to improve healthcare delivery, particularly in underserved contexts. Human-centered design and the Hasso-Plattner Institute were two other design thinking approaches used in the reviewed articles. These approaches are described in the following sections.

4) Human centered design: Innovation, Design Engineering Organization (IDEO) created the human-centered design (HCD) approach. Hearing, Creating, and Delivering are the three phases of HCD[2]. The author in [90] states that "HCD will assist one in hearing the needs of users in new ways, creating innovative solutions to meet users' needs, and delivering answers with financial sustainability in mind." (Page 7) HCD was used in the reviewed articles to create a patient-centered e-health solution for patients receiving weight reduction therapies [50].

5) Hasso-Plattner Institute(HPI): The Hasso-Plattner Institute's approach [3] consists of six steps: 'Understand,' 'Observe,' 'Point of View,' 'Ideate,' 'Prototype,' and 'Test.' In the reviewed articles, HPI was used to improve the testing of latent tuberculosis infection by health workers [75].

6) Techno-centric appraoch: The term "technocentric approach" refers to a point of view that emphasizes technological aspects of designs. The primary distinction between the techno-centric approach and the other approaches described in the reviewed articles is that the other approaches are focused on understanding potential users, which is typically discovered through research process by conducting comprehensive analysis of potential users' behaviours, actions, and desires. The techno-centric approach, on the other hand, is based on "technology-push," in which designers focus on technology first, after which try for implementations for it. [91].

In the reviewed articles, an example of a techno-centric approach was used in design of mHealth to predict patients' ambience, feelings, psychological states, actions, environmental factors, and social context by using machine learning models [58], to improve the design of client adherence to therapeutic activities [59], to collect users' psychological, physiological, and activity information [62][64], and to promote postpartum weight loss [73].

Fig. 7 presents the major design process approaches employed in the design of mHealth interventions for user engagement.

The articles under consideration took more than one approach. Overall, 21 (66%) of the 32 reviewed articles used a participatory design approach, while 8 (25%) used a techno-centric approach, and 3 (9%) used design thinking (whether double diamond, human centered design, or Hasso-Plattner Institute), with that being the least used approach in the articles reviewed, as shown in Fig. 7. This suggests that knowledge about how design thinking can impact mHealth intervention competencies is either still developing or has a minor impact in mHealth designs. Scholars, on the other hand, advocate for more research into how design thinking influences innovation design processes and methods [92].

[1] http://www.designcouncil.org.uk

[2] https://designthinking.ideo.com/

[3] http://www.hpi.uni-potsdam.de/d_school/designthinking

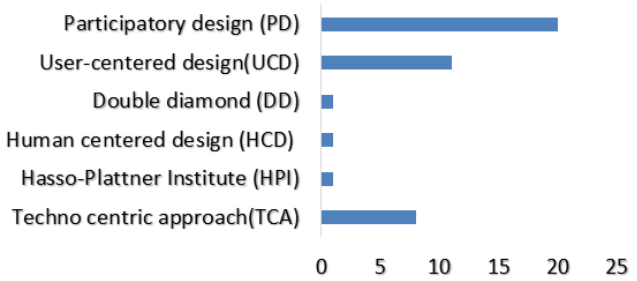


Fig. 7. mHealth Design Approach.

F. Socio-technical Aspects

Socio-technical aspects describe case-specific interventions based on qualitative and empirical evidence and include three descriptive elements such as early user engagement in design, evaluations of mHealth interventions, and understanding users' socio-cultural context.

Eleven studies (34%) used qualitative design method to examine various components of the system, including personal traits of use as well as larger issues of patient-healthcare-system interaction [61]. Examples include early user involvement and identifying and managing relationships between stakeholders in the design of mHealth interventions [71][77], user involvement in the evaluation stages of mHealth development [8][73], and incorporation of users' socio-cultural contexts into the design of mHealth interventions [51] [65]. The sociocultural contexts of mHealth intervention use, according to [49], are among the most difficult aspects of developing mHealth solutions and designing for technology acceptance and adoption. Two of the most important socio-cultural factors to consider before developing any mHealth solution are the position of users in the design of systems and products, as well as cultural differences. Thus, a culture-rooted design approach is considered to be the best way to actually connect and communicate cultural identity, significance, values, and tradition [93]. Of the reviewed articles, 21 (66%) did not explicitly discuss sociotechnical aspects in the design process of mHealth interventions.

G. Design Evaluations

Design evaluations are the processes that are used to determine whether mHealth interventions work as intended for end users.

In the 7 of the 32 reviewed articles (22 %) that discussed evaluation, the following four techniques were used to evaluate design: semi structured interviews, usability inspection, survey, and randomized control trial. Table II summarizes the evaluation techniques, their benefits and drawbacks, and the design goals.

These various evaluation techniques in Table II can be divided into two categories: formative and summative. Formative evaluations, according to [47], are studies that aim to inform the development and design of effective intervention techniques. Summative evaluations, on the other hand, are techniques carried out at the completion of an intervention (or at the end of each stage of the intervention) to verify the degree to which expected results were achieved.

TABLE II. EVALUATION TECHNIQUES AND DESIGN GOALS AS IDENTIFIED IN THE REVIEWED ARTICLES

Articles	Evaluation techniques	Descriptions	Design Goals	Considerations
"Ref. [65]"	Semi structured	This technique provides a channel for the distribution of UXs and sensations in the design process of mobile phone video messaging smoking cessation intervention and multimedia messaging depression prevention intervention.	Empathic design:	Benefits: inform changes to increase satisfaction, interaction, and adapt to end-user needs; identify a variety of issues associated with intervention use Drawbacks: subject to bias, especially if there isn't enough time to collect and transcribe data and it takes more than one person to decide on the themes generating.
"Ref. [79]"	Usability inspection	The goal of this technique is to first, incorporate real-time user experience by delivering a task to users and observing them as they complete the task … and second, refining the content, potential functionality, and interface … based on user feedback.	User experience :	Benefits: allows to determine which features of the intervention influenced user engagement in real time. Drawbacks: may place a mental burden on the users, making it difficult for the observer to analyze the data collected.
"Ref. [8]"	Usability inspection		User experience and Empathic design	
"Ref. [67]"	Randomized controlled trial	A planned experiment that compares the effectiveness of an intervention.	Empathic design process and inclusiveness	Benefits: high standard of study design
"Ref. [73]"	Randomized controlled trial	Randomized controlled, trial to test the efficacy of a SmartPhone-based intervention to promote postpartum weight loss.	Efficacy, predictive and personalized	Drawbacks: Ethical concerns and the difficulty of randomizing subjects

"Ref. [77]"	Survey	Structured questionnaires with many questions to elicit comprehensive information ….	User experience, efficacy and inclusiveness	Benefit: It may be less expensive than alternatives. Drawbacks Does not give an indication of the sequence of events because they are carried out at one time point.
"Ref. [75]"	Usability inspection	Incorporate real-time user experience by delivering a task to users and observing them as they complete the task in the design of a mHealth intervention and analyzing and reporting the results through the lens of the combined Information Systems Research (ISR) framework and design thinking approach.	User experience, efficacy, predictive and personalized and some extent empathic design process	Benefits: allows to determine which features of the intervention influenced user engagement, effective and usable in real time. Drawbacks: may place a mental burden on the users.

H. Checklist of Design Process to Improve user Engagement

We developed a checklist based on a comprehensive evaluation of the 32 articles that considers 6 themes of the design process and the corresponding implementations. There were 16 items reported that improve user engagement in total, and we provide explanations and illustrations as a basis for future research (Table III).

TABLE III. CHECKLIST OF DESIGN PROCESS THAT ENHANCE USER ENGAGEMENT

Themes	**Criteria**
Goal of the design	1. Outline in a clear statement the goals of the design. 2. Identify metrics that will allow mHealth designers and developers to track progress and determine when the design's goals have been met.
Design target populations	3. Identify stakeholders and their roles in the mobile health design. 4. Include stakeholders with a diverse set of skills and perspectives and 5. Involve all stakeholders from the beginning and throughout the design process of mHealth interventions. 6. Examine the ethical issues surrounding participant enrollment, such as obtaining consent and maintaining confidentiality. 7.
Design method	8. Base the design on a clear understanding of the users, tasks performance, and environments. 9. Provide task performance and environment information tailored to the user's preferences.
Design approach	10. Use an iterative design process. 11. Ensure that the design process considers the entire user experience, flow and aesthetics.
Socio-technical aspects	12. Use Techno-social design (TSD) and culture centered design (CCD) to incorporate users' socio-cultural contexts into the design of mHealth interventions. 13. Co-designing with users to understand their values is preferable to designing for users. 14. Identify the barriers and facilitators of intervention participation among study participants. Individual-level structural barriers or facilitators, as well as other factors that may limit a user's ability to engage with the intervention, should be addressed.
Design evaluations	15. Assess how users felt about the intervention or how satisfied they were with it. 16. Describe the evaluation techniques used (for example, usability testing), along with the target group(s). Determine whether the mHealth intervention incurred costs that were proportionate to the benefits of the design's goal.

V. DISCUSSION

In response to several requests for more information on how to design mHealth interventions that effectively engage their users and measure engagement, this systematic literature review paints a picture of how these design processes address user engagement and their socio-cultural contexts. As a methodological framework, user-centered design is used, and related projects that explicitly apply that framework are presented. Based on the articles we reviewed, we used thematic synthesis to identify design processes that increased user engagement with mHealth interventions. This paper discusses six analytical themes related to the design process that can strengthen user engagement with mHealth interventions. We created a design process checklist that improves user engagement to encourage better application of the study's findings to future mHealth intervention development. This tool contains 16 evidence-based items that are clearly described for mHealth intervention designers and developers.

This research yields four major findings. To begin, a robust design process for user engagement with mHealth that incorporates users' socio-cultural contexts into the design of mHealth interventions is required. It has been established that the socio-cultural contexts associated with user engagement are important factors to consider when attempting to achieve user engagement with technology [11]. Techno-social design (TSD) and culture-centered design (CCD) are design principles that emphasize users' social and cultural backgrounds [93]. CCD focuses on the target user and their specific cultural situation. It offers a complementary, rather than diametrically opposed perspective to existing design

methodologies [93]. These design approaches would incorporate user feedback into the mHealth intervention design process. Ethnographic observation or any other methods of describing the mechanism of user engagement with mHealth applications and designing engaging mHealth applications could be used in user research.

Second, the systematic review of literature revealed that mHealth designs should explicitly use user- or human-centered design approaches and involve users from the beginning to understand their needs, as well as throughout the mHealth design process. This finding is consistent with user-centered design interpretations and requirements, which assert that such a process must involve users and understand their needs early and throughout an iterative process.

Third, mHealth initiatives could conduct initial evaluations by soliciting user feedback (for example, through semi-structured interviews or surveys), users interacting with mHealth application prototypes using the usability inspection method and asking detailed questions of what users recognized from the prototype and how it engaged them during user testing.

Finally, more collaboration between patient and clinician populations in mHealth design should be allowed in all mHealth projects. Involvement and engagement of stakeholders (patients and clinicians) in research teams should be encouraged in order to foster an appropriate partnership. To keep the use of mHealth tools relevant, [94] states that a technology-enabled health care partnership with patients and clinicians is required.

We acknowledge some limitations in our work, namely that the literature on mHealth design processes is primarily focused on efficacy and general user experience, as well as other design objectives, with eleven of 32 articles emphasizing the importance of identifying socio-cultural contexts of the user group in the mHealth design process. The articles, however, did not provide extensive and clear guidance, frameworks, or methods for uncovering such socio-cultural contexts in order to improve user engagement with mHealth technology. There was little evidence of these design processes being evaluated (as opposed to mHealth intervention evaluations). As a result, there is a gap in understanding and examining the factors that influence mHealth engagement, acceptance, and usage processes within the mHealth design process. This gap explains why mHealth interventions are poorly accepted and have a limited impact.

Despite these limitations and constraints, we believe the research findings have significant implications, prompting us to make the following recommendations.

1) An ideal mHealth app user engagement framework should capture the users' sociocultural contexts in order to assist mHealth developers and implementers in determining which aspects of the interaction with technology engage, or fail to engage, users. This would overcome the limitations of previous frameworks by covering the design of mHealth apps for user engagement.

2) To assess user engagement with an app, mHealth evaluation criteria should be clear, concise, specific, and objective. It is also critical to assess user engagement early in the design process of mHealth interventions and consider improving the user-centered design framework and perhaps using other frameworks, particularly techno-social design (TSD) and cultural centered design (CCD), that consider design principles emphasizing users' social and cultural contexts in the design of mHealth interventions.

3) A comprehensive objective mHealth user engagement design framework requires future testing on various platforms across many mHealth implementations to determine a low-burden approach to improve user engagement cheaply and efficiently.

4) There are contributions about theory-based mHealth systems for, user-engaged mHealth interventions based on behavioural techniques. However, the reviews emphasize the need to develop a framework that will employ processes and tools that uncover socio-cultural contexts of end users into the design of mHealth technologies and encompasses all the areas that this systematic review identified as important for user engagement with mHealth.

VI. Conclusion

This systematic review of the literature looks at articles that focus on the design processes of mHealth interventions. We provided a thorough comparison of the design methods and who the design targets for mHealth design interventions, highlight trends in the mobile design process, targeting patients and/or clinicians, including design goals implemented alongside the mHealth interventions, design process approach used, sociotechnical aspects of the systems, and mHealth intervention evaluations. The strengths and weaknesses of existing mHealth design processes for user engagement are discussed, and recommendations for future research in these areas are made. We discovered that only a few of the reviewed articles considered the evaluation of mHealth interventions, and the majority of the articles did not consider a framework that will incorporate processes and tools that uncover end users' socio-cultural contexts into the design of mHealth technologies. According to the findings, the participatory approach of user-centered designs was most frequently used in the review articles.

References

[1] R. Heeks. Information systems and developing countries: Failure, success, and local improvisations. The information society, 18(2), 2002, pp. 101-112.

[2] P. Bhatt, A.J. Ahmad, & M. A. Roomi. Social innovation with open source software: User engagement and development challenges in India. Technovation, 52, 2016, pp. 28-39.

[3] A. K. Böhm, M. L. Jensen, M. R. Sørensen, & T. Stargardt, . Real-world evidence of user engagement with mobile health for diabetes management: longitudinal observational study. JMIR mHealth and uHealth, 8(11), 2020. e22212.

[4] A. Grady, S. Yoong, R. Sutherland, H. Lee, N. Nathan, & L. Wolfenden, L. Improving the public health impact of eHealth and mHealth interventions. Australian and New Zealand journal of public health, 42(2), 2018. pp.118-119.

[5] M. Cherubini, & N. Oliver,. A refined experience sampling method to capture mobile user experience. 2009. arXiv preprint arXiv:0906.4125.

[6] L.A. Kappelman, & E. R. McLean. User engagement in the development, implementation, and use of information technologies. In HICSS (4) 1994. pp. 512-521.

[7] S. Attfield, G. Kazai, M. Lalmas, & B. Piwowarski. Towards a science of user engagement (position paper). In WSDM workshop on user modelling for Web applications 2011. pp. 9-12.

[8] R. Schnall, M. Rojas, S. Bakken, W. Brown, A. Carballo-Dieguez, M. Carry, ... & J. Travers, J. A user-centered model for designing consumer mobile health (mHealth) applications (apps). Journal of biomedical informatics, 60, 2016. Pp. 243-251.

[9] K.E. Curtis, S. Lahiri, & K.E. Brown. Targeting parents for childhood weight management: development of a theory-driven and user-centered healthy eating app. JMIR mHealth and uHealth, 3(2), 2015. e3857.

[10] F. Verhoeven, K. Tanja-Dijkstra, N. Nijland, G. Eysenbach, & van L. Gemert-Pijnen, . Asynchronous and synchronous teleconsultation for diabetes care: a systematic literature review. Journal of diabetes science and technology, 4(3), 2010. pp. 666-684.

[11] T. McCurdie, S. Taneva, M. Casselman, M. Yeung, C. McDaniel, W. Ho, & J. Cafazzo. mHealth consumer apps: the case for user-centered design. Biomedical instrumentation & technology, 46(2), 2012. 49.

[12] P. CISION. Motivating Patients to Use Smartphone Health Apps. 2015. URL: http://www. prweb. com/releases/2011/04/prweb5268884. htm [accessed 2015-08-10].

[13] G.C. Machado, M.B. Pinheiro, H. Lee, O.H. Ahmed, P. Hendrick, C. Williams, & S.J. Kamper. Smartphone apps for the self-management of low back pain: A systematic review. Best Practice & Research Clinical Rheumatology, 30(6), 2016. Pp.1098-1109.

[14] S. Taki, S. Lymer, C.G. Russell, K. Campbell, R. Laws, K.L. Ong, ... & E. Denney-Wilson, E. Assessing user engagement of an mHealth intervention: development and implementation of the growing healthy app engagement index. JMIR mHealth and uHealth, 5(6), 2017. e7236.

[15] O. Perski, D. Crane, E. Beard, & J. Brown. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digital health, 5, 2019. 2055207619880676.

[16] N. Wickramasinghe. Understanding the mHealth implementation and adoption impact: a FVM perspective. Electron. Sci. Technol. Appl, 5(2), 2018. p.1-9.

[17] N.A. Shozi, D. Pottas, & N. Mostert-Phipps. A socio-technical perspective on the use of mobile phones for remote data collection in home community based care in developing countries. In International Conference on e-Infrastructure and e-Services for Developing Countries 2011. pp. 135-145. Springer, Berlin, Heidelberg.

[18] K. Tabi, A.S. Randhawa, F. Choi, Z. Mithani, F. Albers, M. Schnieder, ... & M. Krausz. Mobile apps for medication management: review and analysis. JMIR mHealth and uHealth, 7(9), 2019. e13608.

[19] C. Avgerou, & G. Walsham. Information technology in context: Implementing systems in the developing world. 2000. Brookfield, VT: Ashgate Publishing.

[20] M. R Hoque, M.S. Rahman, N.J. Nipa, & M.R. Hasan. Mobile health interventions in developing countries: A systematic review. Health Informatics Journal, 2020. 1460458220937102.

[21] D. Zurovac, A.O. Talisuna, & R.W. Snow. Mobile phone text messaging: tool for malaria control in Africa. PLoS medicine, 9(2), 2012. e1001176.

[22] C.B. Aranda-Jan, N. Mohutsiwa-Dibe, & S. Loukanova, S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC public health, 14(1), 2014. pp.1-15.

[23] E.F. Krah, & J.G. de Kruijf, Exploring the ambivalent evidence base of mobile health (mHealth): A systematic literature review on the use of mobile phones for the improvement of community health in Africa. Digital health, 2, 2016. 2055207616679264.

[24] C.S. Hall, E. Fottrell, S. Wilkinson & P. Byass. Assessing the impact of mHealth interventions in low-and middle-income countries–what has been shown to work?. Global health action, 7(1), 2014. 25606.

[25] C. Granja, W. Janssen & M.A. Johansen. Factors determining the success and failure of eHealth interventions: systematic review of the literature. Journal of medical Internet research, 20(5), 2018. e10235.

[26] T.D. Manda & Y. Msosa,. Socio-technical arrangements for mhealth: Extending the mobile device use and adoption framework. In International Conference on e-Infrastructure and e-Services for Developing Countries. 2012. pp. 208-217. Springer, Berlin, Heidelberg.

[27] J. G. Kahn, J.S. Yang, & J.S. Kahn. 'Mobile'health needs and opportunities in developing countries. Health affairs, 29(2), 2010. pp.252-258.

[28] J. Kaasbøll, & J. L. Nhampossa. Transfer of public sector information systems between developing countries: south-south cooperation. Social Implications of Computers in Developing Countries, Bangalore. 2002.

[29] J. H. Wu, S.C. Wang & L.M. Lin. Mobile computing acceptance factors in the healthcare industry: A structural equation model. International journal of medical informatics, 76(1), 2007. pp.66-77.

[30] E.B. Tate, D. Spruijt-Metz, G. O'Reilly, M. Jordan-Marsh, M. Gotsis, M.A. Pentz, & G. F. Dunton. mHealth approaches to child obesity prevention: successes, unique challenges, and next directions. Translational behavioral medicine, 3(4), 2013. pp.406-415.

[31] R. Harris, & R. Davison. Anxiety and involvement: Cultural dimensions of attitudes toward computers in developing societies. In Global perspective of information technology management 2002. pp. 234-259. IGI Global.

[32] G. Hofstede, G. Dimensionalizing cultures: The Hofstede model in context. Online readings in psychology and culture, 2(1), 2011. pp.307-0919.

[33] J. Braa, O. Hanseth, A. Heywood, W. Mohammed, & V. Shaw. Developing health information systems in developing countries: the flexible standards strategy. Mis Quarterly, 2007. pp.381-402.

[34] T. Ikwunne, L. Hederman, & P.J Wall. Designing Mobile Health for User Engagement: The Importance of Socio-Technical Approach. 2021. arXiv preprint arXiv:2108.09786.

[35] R. Braun, C. Catalani, J. Wimbush, & D. Israelski, D. Community health workers and mobile technology: a systematic review of the literature. PloS one, 8(6), 2013. e65772.

[36] K. Overbeeke, T. Djajadiningrat, C. Hummels, S. Wensveen, & J. Prens. Let's make things engaging. In Funology . 2003.pp. 7-17. Springer, Dordrecht.

[37] H. L. O'Brien, & E.G. Toms. The development and evaluation of a survey to measure user engagement. Journal of the American Society for Information Science and Technology, 61(1), 2010. pp.50-69.

[38] K. Doherty, K & G. Doherty. Engagement in HCI: conception, theory and measurement. ACM Computing Surveys (CSUR), 51(5),2018. pp. 1-39.

[39] B. Laurel. Computers as theatre reading. Mas: Addison-Wesley Publishing Company. 1991.

[40] W. Quesenbery. Dimensions of usability. Content and complexity: Information design in technical communication. 2003.

[41] P. Saariluoma. Explanatory frameworks for interaction design. In. A. Pirhonen, P. Saariluoma, H Isomaki, & C. Roast (Eds.) Future interaction design 2005.pp. 67-83.

[42] H.L. O'brien, & E.G. Toms. Examining the generalizability of the User Engagement Scale (UES) in exploratory search. Information Processing & Management, 49(5), 2013. pp.1092-1107.

[43] J. McCarthy, & P. Wright. Technology as experience. interactions, 11(5), 2004. pp.42-43.

[44] S. J. Flaherty, M. McCarthy, A.M. Collins, C. McCafferty, & F. M. McAuliffe. Exploring engagement with health apps: the emerging importance of situational involvement and individual characteristics. European Journal of Marketing.2021.

[45] V. Braun & V. Clarke. Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 2006. pp.77–101. Available at: http://eprints.uwe.ac.uk/11735.

[46] L.L. Novak, R. J. Holden, S. H. Anders, J.Y. Hong, & B.T. Karsh, B. Using a sociotechnical framework to understand adaptations in health IT implementation. International journal of medical informatics, 82(12), 2013. e331-e344.

[47] World Health Organization. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment.2016.

[48] K. Best. Design management: managing design strategy, process and implementation. AVA publishing. 2006.

[49] B. Aryana, L. Brewster & J.A. Nocera. Design for mobile mental health: an exploratory review. Health and Technology, 9(4), 2019. pp.401-424.

[50] A. Das, & D. Svanæs,. Human-centred methods in the design of an e-health solution for patients undergoing weight loss treatment. International journal of medical informatics, 82(11), 2013. pp.1075-1091.

[51] G. Marcu, J.E. Bardram, & S. Gabrielli. A framework for overcoming challenges in designing persuasive monitoring and feedback systems for mental illness. In 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops 2011.pp. 1-8. IEEE.

[52] J. E. Bardram, M. Frost, K. Szántó, M. Faurholt-Jepsen, M. Vinberg, & L.V. Kessing. Designing mobile health technology for bipolar disorder: a field trial of the monarca system. In Proceedings of the SIGCHI conference on human factors in computing systems 2013. pp. 2627-2636.

[53] A. K. Bright & L. Coventry, L. Assistive technology for older adults: psychological and socio-emotional design requirements. In Proceedings of the 6th International Conference on Pervasive Technologies Related to Assistive Environments 2013.pp. 1-4.

[54] M. Price, E.K Yuen, E.M. Goetter, J.D. Herbert, E.M. Forman, R. Acierno, and K.J Ruggiero. mHealth: a mechanism to deliver more accessible, more effective mental health care. Clinical psychology & psychotherapy, 21(5), 2014. pp.427-436.

[55] D.D. Luxton, R.A. McCann, N.E. Bush, M.C. Mishkind, and G.M. Reger. mHealth for mental health: Integrating smartphone technology in behavioral healthcare. Professional Psychology: Research and Practice, 42(6), 2011. p.505.

[56] A. Gaggioli, G. Pioggia, G. Tartarisco, G. Baldus, D. Corda, P. Cipresso, & G. Riva. A mobile data collection platform for mental health research. Personal and Ubiquitous Computing, 17(2), 2013. pp. 241-251.

[57] J.F. Pelletier, M. Rowe, N. François, J. Bordeleau, & S. Lupien. No personalization without participation: on the active contribution of psychiatric patients to the development of a mobile application for mental health. BMC Medical Informatics and Decision Making, 13(1), 2013. pp.1-8.

[58] M.N. Burns, M. Begale, J. Duffecy, D. Gergle, C.J. Karr, E. Giangrande, & D.C. Mohr. Harnessing context sensing to develop a mobile intervention for depression. Journal of medical Internet research, 13(3), 2011. e55.

[59] M. Matthews & G. Doherty. In the mood: engaging teenagers in psychotherapy using mobile phones. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2011.pp. 2947-2956.

[60] S. C. Reid, S.D. Kauer, S.J. Hearps, A.H. Crooke, A.S. Khor, L.A Sanci, & G.C.Patton. A mobile phone application for the assessment and management of youth mental health problems in primary care: a randomised controlled trial. BMC family practice, 12(1), 2011. pp.1-14.

[61] J.N. Stinson, C. Lalloo, L. Harris, L. Isaac, F. Campbell, S. Brown, ... & A. Karim. iCanCope with Pain™: User-centred design of a web-and mobile-based self-management program for youth with chronic pain based on identified health care needs. Pain Research and Management, 19(5), 2014. pp.257-265.

[62] Y.X. Chen, Y.P. Hung & H.C. Chen. Mobile application–Assisted cognitive behavioral therapy for insomnia in an older adult. Telemedicine and e-Health, 22(4), 2016. pp.332-334.

[63] J. Huh, C.J. Cerrada, E. Dzubur, G.F. Dunton, D. Spruijt-Metz, & A.M. Leventhal. Effect of a mobile just-in-time implementation intention intervention on momentary smoking lapses in smoking cessation attempts among Asian American young adults. Translational behavioral medicine, 11(1), 2021. 216-225.

[64] N. Bidargaddi, G. Schrader, P. Klasnja, J. Licinio, S. & Murphy. Designing m-Health interventions for precision mental health support. Translational psychiatry, 10(1), 2020. pp.1-8.

[65] R. Whittaker, S. Merry, E. Dorey & R. Maddison. A development and evaluation process for mHealth interventions: examples from New Zealand. Journal of health communication, 17(sup1), 2012. pp.11-21.

[66] T.A. Dennis & L. J. O'Toole. Mental health on the go: Effects of a gamified attention-bias modification mobile application in trait-anxious adults. Clinical Psychological Science, 2(5), 2014. pp.576-590.

[67] H. Lee, R. Ghebre, C. Le, Y.J. Jang, M. Sharratt, & D. Yee. Mobile phone multilevel and multimedia messaging intervention for breast cancer screening: pilot randomized controlled trial. JMIR mHealth and uHealth, 5(11), 2017. e154.

[68] M. Dugas, K. Crowley, G.G. Gao, T. Xu, R. Agarwal, A.W. Kruglanski, & N. Steinle,. Individual differences in regulatory mode moderate the effectiveness of a pilot mHealth trial for diabetes management among older veterans. PloS one, 13(3), 2018. e0192807.

[69] J.K Carroll, J.N. Tobin, A. Luque, S. Farah, M. Sanders, A. Cassells, ... & K. Fiscella. "Get ready and empowered about treatment"(GREAT) study: a pragmatic randomized controlled trial of activation in persons living with HIV. Journal of general internal medicine, 34(9), 2019. pp.1782-1789.

[70] P. Mc Kenna, G. Babughirana, M. Amponsah, S.G. Egoeh, E. Banura, R. Kanwagi, & B. Gray. Mobile training and support (MOTS) service—using technology to increase Ebola preparedness of remotely-located community health workers (CHWs) in Sierra Leone. Mhealth, 5. 2019.

[71] D. Ben-Zeev, C.J. Brenner, M. Begale, J. Duffecy, D.C. Mohr & K.T. Mueser. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. Schizophrenia bulletin, 40(6), 2014. pp.1244-1253.

[72] G. T. Ahsan, I.D Addo, S.I. Ahamed, D. Petereit, S. Kanekar, L. Burhansstipanov, L.U & Krebs. Toward an mHealth intervention for smoking cessation. In 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops 2013. pp. 345-350. IEEE.

[73] L.A. Gilmore, M.C. Klempel, C.K. Martin, C.A. Myers, J.H. Burton, E. F. Sutton, & L.M. Redman. Personalized mobile health intervention for health and weight loss in postpartum women receiving women, infants, and children benefit: a randomized controlled pilot study. Journal of Women's Health, 26(7), 2017. pp.719-727.

[74] U. Müssener, M. Löf, P. Bendtsen, & M. Bendtsen. Using mobile devices to deliver lifestyle interventions targeting at-risk high school students: protocol for a participatory design study. JMIR research protocols, 9(1), 2020. e14588.

[75] J. Farao, B. Malila, N. Conrad, T. Mutsvangwa, M.X. Rangaka, & T.S.Douglas). A user-centred design framework for mHealth. PloS one, 15(8), 2020. e0237910.

[76] A.E. Kazdin, & S.M. Rabbitt. Novel models for delivering mental health services and reducing the burdens of mental illness. Clinical Psychological Science, 1(2), 2013. pp.170-191.

[77] D. R. Nyatuka, & R. de la Harpe. Evaluating mHealth Interventions in an Underserved Context Using Service Design Strategy: A Case of Kenya. In Proceedings of the third International Conference on Medical and Health Informatics 2019. pp. 153-160.

[78] S.P. Jones, V. Patel, S. Saxena, N. Radcliffe, S. Ali Al-Marri, & A. Darzi. How Google's 'ten things we know to be true'could guide the development of mental health mobile apps. Health Affairs, 33(9), 2014. pp. 1603-1611.

[79] R. Schnall, M. Rojas, J. Travers, W Brown III, & S. Bakken. Use of design science for informing the development of a mobile app for persons living with HIV. In AMIA Annual Symposium Proceedings 2014, p. 1037. American Medical Informatics Association.

[80] H. D. Nguyen, D.C.C Poo, H. Zhang, & W. Wang. Analysis and design of an mHealth intervention for community-based health education: an empirical evidence of coronary heart disease prevention program among working adults. In International Conference on Design Science Research in Information System and Technology 2017.pp. 57-72. Springer, Cham.

[81] V. P. Cornet, T. Toscos, D. Bolchini, R.R Ghahari, R. Ahmed, C. Daley, ... & R.J Holden . Untold stories in user-centered design of mobile health: Practical challenges and strategies learned from the design and

evaluation of an app for older adults with heart failure. JMIR mHealth and uHealth, 8(7), 2020.e17703.

[82] A.M. Polhemus, J. Novák, J. Ferrao, S. Simblett, M. Radaelli, P. Locatelli, ... & M. Hotopf. Human-centered design strategies for device selection in mhealth programs: development of a novel framework and case study. JMIR mHealth and uHealth, 8(5), 2020. e16043.

[83] E. B. N. Sanders. From user-centered to participatory design approaches. In Design and the social sciences 2002. pp. 18-25. CRC Press.

[84] J. Ross, F. Stevenson, R. Lau, and E. Murray. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implementation science, 11(1), 2016. p.146.

[85] P. Vink, A.S. Imada, & K.J. Zink. Defining stakeholder involvement in participatory design processes. Applied ergonomics, 39(4), 2008. p.519-526.

[86] ISO. Ergonomics of Human-system Interaction: Part 210: Human-centred Design for Interactive Systems. 2010. ISO.

[87] ISO Human-centred design processes for interactive systems. Geneva: 1999.ISO.

[88] Y.J. Korpershoek, S. Hermsen, L. Schoonhoven, M.J. Schuurmans, & J.C. Trappenburg. User-centered design of a mobile health intervention to enhance exacerbation-related self-management in patients with chronic obstructive pulmonary disease (copilot): mixed methods study. Journal of medical Internet research, 22(6), 2020. e15449.

[89] A. Good & O. Omisade. Linking Activity Theory with User Centred Design: A Human Computer Interaction Framework for the Design and Evaluation of. Applied Interdisciplinary Theory in Health Informatics: A Knowledge Base for Practitioners, 2019. p263, 49.

[90] K. Tschimmel. Design Thinking as an effective Toolkit for Innovation. In ISPIM Conference Proceedings 2012. p. 1. The International Society for Professional Innovation Management (ISPIM).

[91] E.G. Carayannis, L. Dezi, G. Gregori, & E. Calo . Smart environments and techno-centric and human-centric innovations for Industry and Society 5.0: A Quintuple Helix Innovation System view towards smart, sustainable, and inclusive solutions. Journal of the Knowledge Economy, 2021. pp.1-30.

[92] P. Micheli, S. J. Wilner, S. H. Bhatti, M. Mura, & M.B. Beverland. Doing design thinking: Conceptual review, synthesis, and research agenda. Journal of Product Innovation Management, 36(2), 2019. pp.124-148.

[93] S.T. Shen, M. Woolley & S. Prior. Towards culture-centred design. Interacting with computers, 18(4), 2006. pp.820-852.

[94] L. Opipari-Arrigan, D.M. Dykes, S.A. Saeed, S. Thakkar, L. Burns, B.A. Chini, ... & H.C. Kaplan. Technology-Enabled Health Care Collaboration in Pediatric Chronic Illness: Pre-Post Interventional Study for Feasibility, Acceptability, and Clinical Impact of an Electronic Health Record–Linked Platform for Patient-Clinician Partnership. JMIR mHealth and uHealth, 8(11), 2020. e11968.

An Intelligent Metaheuristic Optimization with Deep Convolutional Recurrent Neural Network Enabled Sarcasm Detection and Classification Model

K.Kavitha[1]
Research Scholar
Department of Computer Science and Engg
Acharya Nagarjuna University, Guntur
Assistant Professor, AITAM, Tekkali

Suneetha Chittieni[2]
Associate Professor
Department of Computer Applications
RVR & JC College of Engineering
Guntur, India

***Abstract*—Sarcasm is a state of speech in which the speaker says something that is externally unfriendly with a purpose of abusing/deriding the listener and/or a third person. Since sarcasm detection is mainly based on the context of utterances or sentences, it is hard to design a model to proficiently detect sarcasm in the domain of natural language processing (NLP). Despite the fact that various methods for detecting sarcasm have been created utilizing statistical machine learning and rule-based approaches, they are unable of discerning figurative meanings of words. The models developed using deep learning approaches have shown superior performance for sarcasm detection over traditional approaches. With this motivation, this paper develops novel deep learning (DL) enabled sarcasm detection and classification (DLE-SDC) model. The DLE-SDC technique primarily involves pre-processing stage which encompasses single character removal, multispaces removal, URL removal, stop word removal, and tokenization. Next to data preprocessing, the preprocessed data is converted into the feature vector by Glove Embeddings technique. Followed by, convolutional neural network with recurrent neural network (CNN-RNN) technique is utilized to detect and classify sarcasm. In order to boost the detection outcomes of the CNN+RNN technique, a hyper parameter tuning process utilizing teaching and learning based optimization (TLBO) algorithm is employed in such a way that the classification performance gets increased. The DLE-SDC model is validated using the benchmark dataset and the performance is examined interms of precision, recall, accuracy, and F1-score.**

Keywords*—Sarcasm detection; data classification; deep learning; feature extraction; TLBO algorithm; parameter optimization

I. INTRODUCTION

Sarcasm detection in discussions has become ever more popular amongst natural language processing (NLP) scientists with the greater usage of communicative threats on social networking platforms. Natural language is an essential data source of human emotions. Automatic sarcasm detection is repeatedly defined as an NLP problem since it mainly needs to understand the human emotions language, expressions expressed by the non-textual/textual content. Sarcasm detection has gained more attention in previous decades since it facilitates precise analysis in online reviews and comments [1]. As an illustrative approach, sarcasm utilizes word in a manner which differs from the traditional meaning and order as result of misleading polarity classification. The results obtained in this development can be used for information categorization.

Sarcasm could be deliberated as an implied form of emotion. Usually, it transmits the reverse of what has been aimed. Generally, Sarcasm is related to literary devices like satire and wit/irony i.e., utilized for insult, refutes, amuse or make fun of. Specifically, the teacher exclaimed "Credit to your hard work. I have been never impressed more in my lifetime. Lol!" these sentences might expose i.e., gratitude. But, the expression of a speaker and context demonstrate the sarcastic manner of these expressions. In the lack of visible expression, defining sarcasm in Twitter is a challenging one. A stimulating perception of sarcasm has been proposed by [2] in which the analyses were carried out in 2 sarcastic states: all centric and egocentric. The previous terms indicate that the sarcasm was observed/felt only from the participant's point of view and not from addressees' perception and the last one indicates sarcasm being observed from the addressee and participant perspectives. The generic understanding of the result transmits the prosodic feature, the one including pattern of sounds and stress is more useful in identifying sarcasm when compared to contextual features.

Fundamental analyses of sentiment from the text mightn't be effective for understanding the clear stimulation because of the existence of different literary devices like irony, sarcasm, and so on [3]. Thus, sarcasm detection is highly required for avoiding all kinds of misinterpretation in all kinds of transmission and for ensuring that meaning aimed in the statement is assumed accordingly. Automatically identifying sarcasm could be a difficult task that could be demonstrated by automatic sarcasm analysis and detection. Identifying sarcastic statements becomes an essential process in social networking applications since it effects the organization that mines social networking data. In spite of the existence of several potential features are extracted from text, they could be gathered into major classes, such as contextual, lexical, pragmatic, and hyperbolic features [4]. The fundamental objective of this study is to classify sarcasm into different kinds that aid in understanding the intent to hurt or level of hurt i.e., existing in the sarcastic statements. Because sarcasm may elicit a broad range of feelings in a person, it can either make the receiver

laugh or, in the worst-case it might elicit a deeper sense of emotional harm. The applications of type detection might be effective in understanding the sentiments behindhand sarcasm, which offer a perspective to the sentimental condition of the person engaging in a sarcastic discussion, namely, the one on whom sarcasm was meant and the person who employs sarcasm.

Several machine learning, rule-based, deep learning, and statistical based methods have been stated in related works on automated sarcasm detection in one sentence i.e., frequently based on the content of words in isolation. This involves a variety of methods like multimodal (text image) content [5] sense disambiguation and polarity flip detection in text [6]. Previous research on detecting sarcasm in text includes pragmatic (context) and lexical (content) clues [7] such as sentiments, interjections, and punctuation alterations, which are major indicators of sarcasm [8]. The characteristics in this study are handmade and cannot be generalized due to the presence of metaphorical slang and informal language, which are often used in online communication. Current research [9, 10] use NN for learning contextual and lexical features, eliminating the necessity for handmade features with the development DL method. In this paper, word embeddings are used to train recurrent, deep convolutional, or attention-based neural networks to achieve advanced results on a variety of large-scale datasets.

This paper develops novel deep learning (DL) enabled sarcasm detection and classification (DLE-SDC) model. The DLE-SDC technique primarily involves pre-processing stage which takes place at different levels. Then, Glove Embedding technique is used for the representation of word vectors. Moreover, convolutional neural network with recurrent neural network (CNN-RNN) technique is utilized to detect and classify sarcasm. In order to boost the detection outcomes of the CNN+RNN technique, a hyper parameter tuning process using teaching and learning based optimization (TLBO) algorithm is employed in such a way that the classification performance gets increased. A wide range of simulations take place on benchmark datasets and validate the results interms of different measures.

II. Literature Review

In Nayel et al. [11], a method that relied on a supervised ML approach named SVM was utilized for detecting sarcasm. The presented method was calculated by an ArSarcasm-v2 dataset. The efficiency of the presented method was related to another method provided to sarcasm detection shared task and sentiment analyses. Kumar and Harish [12] proposed a new method for classifying sarcastic text with content based FS technique. The projected method is composed of 2 phase FS methods for selecting better representation features. In initial phase, traditional FS approaches like MI, IG and Chi-square are utilized for selecting appropriate features subset. The selected feature subset is additionally developed by the next phase. In following phase, k-means clustering process is utilized for selecting better representation features between same features. The selected features are categorized by 2 SVM and RF classifiers. Chatterjee et al. [13] designed features to detect sarcasm by realistic features which considered the context of word. The method is depending upon a linguistic method which defines how human differentiate among various kinds of untruth. Later, they train different ML based classifiers and relate their accuracy.

Razali et al. [14] focus on detecting sarcasm in tweets by combining DL derived features with contextual constructed feature sets. A feature set is retrieved from a CNN framework and carefully combined with the handmade feature set. Those custom feature sets are developed based on their contextual explanation. Every feature set is specifically designed for the solitary task of detecting sarcasm. The aim is to find the optimum features. Few sets are beneficial for working even if it is utilized individually. Other sets aren't really substantial without integration. The result of the experiment shows positive based on Precision, Accuracy, F1-measure, and Recall. The integration of features is categorized by ML methods for the purposes of comparison. The LR approach is considered as an optimal classification approach for this work. In Rajeswari and ShanthiBala [15], a supervised classification method viz., MNNB is utilized for detecting sarcasm, and SVM is utilized for detecting the types of sarcasm. In this work, the sarcasm is extracted from the twitters using MNNB. The tweets contain noisy messages and are managed well for efficient detection of sarcasm. Additionally, the types of sarcasm are also detected for diagnosing the state of the user.

Zhang et al. [16] proposed the utilization of NN for detecting sarcasm tweets and compared the impacts of continuous automated features using discrete manual features. Particularly, they utilize bi-directional gated RNN for capturing syntactic and semantic data on twitters, and a pooling NN for extracting contextual features manually from past twitters. Akula and Garibay [17] concentrate on identifying sarcasm in textual conversation from different societal and online platforms for networking. Eventually, they developed an interpretable DL method with gated recurrent units and multi-head self-attention. The major goal of this work [18] is the sentiment analyses of people's opinions exposed on Face book based on the present epidemic condition in lower resource language. To perform this, they have made a large scale dataset consist of 10,742 automatically categorized commentaries in the Albanian language. Moreover, in this study, they reported the effort on the development and design of sentiment analyses based on DL approach. Consequently, they reported the investigational finding attained from this presented sentiment analysis by different classification methods using static and contextualized word embedding, i.e., BERT and fast Text, validated and trained on these curate and collected datasets. Das and Kolya [19] ,the sarcastic word distribution properties of a common pop culture sarcasm corpus, which includes sarcastic speeches and dialogues, are automatically extracted. Further, they proposed an amalgamation of 4p LSTM, each contains unique activation classifier. Those models are mainly intended to effectively identifying sarcasm from the text corpus.

Sundararajan and Palanisamy [20] aim are to enhance the present methods by integrating a novel perception that categorizes the sarcasm on the basis of the levels of harshness applied. The main application of the projected study will be associating the mood of an individual to the types of sarcasm

shown by him/her that can give main perceptions regarding the emotional behaviour of an individual. An ensemble-based FS approach was proposed for choosing the optimum collection of features for detecting sarcasm in tweets. This optimal collection of attributes was used to determine whether the tweets were sarcastic or not. Afterward identifying the sarcastic sentence, a multi-rule based method was projected for determining the sarcasm types. Kumar et al. [21] used Mustard, a typical conversation dataset to determine the use of an ensemble supervised learning approach for identifying sarcasm. Furthermore, it can be useful in reducing model bias and assisting decision makers in knowing how to use this model accurately. Liyuan Liu et al. [22] Proposed a method called A2Text-Net which combines auxiliary variables to improve the performance of sarcastic sentiment classification.

III. The Proposed Model

This study has developed a DLE-SDC technique to classify the presence of sarcasm. The working process is demonstrated in Fig. 1. The proposed method involves different processes namely, preprocessing, Glove based word vector representation, CNN-RNN based classification, and TLBO based parameter optimization.

A. Data Pre-processing

At the first stage, the data is pre-processed to transform into a compatible format. The different sub processes involved in data pre-processing are:

- Remove single letter words.
- Remove multiple spaces.
- Remove punctuation marks.
- Remove numbers.
- Remove stop words and.
- Convert uppercase characters into lowercase.

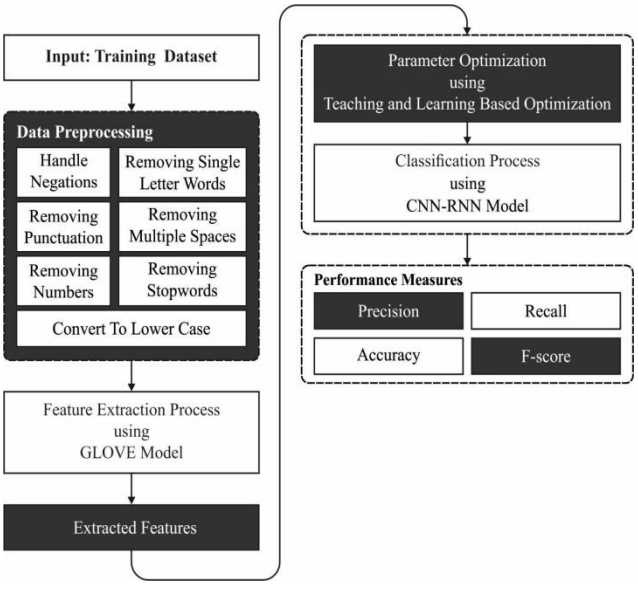


Fig. 1. Working Process of DLE-SDC Model.

B. Glove based Word Representation

The Glove approach can able to generate a vector depiction of words in the application of similarity between words as invariant. It utilizes 2 different methods as CBOW and Skip-gram. The problem related to the conventional methods includes minimum accuracy, maximum processing time, etc. The primary objective of Glove is to incorporate the approaches proposed by 2 techniques wherein optimal accuracy must be assured. In previous to generating Glove approach, the vector depiction of words has been determined. The approaches are employed to generate a vector using standard dimensions (d) for all the words. The approach that employs similarity between 2 words as invariant wherein the words in similar contents is taken into account and show same meaning.

Assume the terms beforehand presenting a formulation of Glove:

- Take a matrix of word to word co-existence count as denoted by X, whereas values X_{ij} store the amount of iterations in a word j in a sentence of word i.
- Assume that $X_i = \sum_k X_{ik}$ represents the amount of times a word could be repeating in content of word i.
- Finally, consider $P_{ij} = P(j|i) = X_{ij}/X_i$ denotes a likelihood of word j displayed in context of word i.

Let us take 2 words i & j i.e., associated with each other in content; e.g., suppose that cricket is a subject matter so that i = duck and j = boundary. Analyzing a ratio of coexistence probability with distinct probe words, k, reveals the relationships between those words. In wordsk, i.e., associated with duck by not including the boundary, let k = out, thus the ratio $P_{ik}/P_{jk}$is maximalized. Similarly, in words k depends on boundary by not including duck, let k = six, so that ratio is minimalized. Hence, words k like score i.e., appropriate to duck and boundary, as ratio is nearly 1. Conventional logic suggests that the proportion of coexistence possibilities might be used as a starting point to calculate the similarity between those terms. The ratio $P_{ik}P_{jk}$ is based on 3 words j, and k, in which standard method simulates the process as follows,

$$F(w_i, w_j, \hat{w}_k) = \frac{P_{ik}}{P_{jk}}, \quad (1)$$

Whereas $w \in R^d$ denotes word vector and $\hat{w} \in R^d$ represents separate context word vector. As vector spaces are integrally linear structures and assume vector variances.

$$F((w_i - w_j)^T, \hat{w}_k) = \frac{P_{ik}}{P_{jk}} \quad (2)$$

The application of algebraic function and group theory are given below:

$$F((w_i - w_j)^T \hat{w}_k) = \frac{F(w_i^T \hat{w}_k)}{F(w_j^T \hat{w}_k)}, \quad (3)$$

$$where, F(w_i^T \hat{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$$

$$w_i^T \hat{w}_k = log(P_{ik}) = \log(X_{ik}) - \log(X_i)$$

$$\exp\left(w_i^T . \widehat{w}_k\right) = P_{ik} = \frac{X_{ik}}{X_i}$$

In which '·' represents a dot product between 2 vectors w_i and $w_k$whereas exp indicates an exponent function.

$$w_i^T . \widehat{w}_k = log(P_{ik}) = log(X_{ik}) - log(X_i), \quad (4)$$

and the $\log(X)$ represents a constant in word i, the aforementioned operation is altered by:

$$w_i^T . \widehat{w}_k + b_i + \hat{\mathrm{b}}_{\mathrm{k}} = log(X_{ik}), \quad (5)$$

In which $b_i$represents a bias for word i and $\hat{\mathrm{b}}_{\mathrm{k}}$indicates a bias for word k.

Later, the optimum function from ML perception is written as follows: RHS from previous equations are calculated from corpus, which should be updated from LHS to produce relevant RHS. Thus, hypothesis (h) determines LHS, while RHS is referred to as output (y). The cost function is then converted using the least square method:

$$J = \sum_{i,k}\left(w_i^T . \widehat{w}_k + b_i + \hat{\mathrm{b}}_{\mathrm{k}} - \log\left(X_{ik}\right)\right)^2, \quad (6)$$

and it is necessary to minimize the cost function. But, previous to employing GD, the cost function should be increase dramatically by some weights to each two words; hence, a cost function may be thought of as a memory that preserves data depending on previously calculated values.

$$J = \sum f\left(X_{ij}\right)\left(w_i^T . \widehat{w}_k + b_i + b_k^{\wedge} - \log\left(X_{ik}\right)\right)^2, \quad (7)$$

whereas $f(X_{ij})$ denotes weight associated with co-existence of term i with j. Generally, f is called as:

$$\left(\frac{x}{x_{\max}}\right)^{\alpha}, if x < x_{\max},$$

$$1, or\ else$$

Later, when a partial derivative of J is handled by w_i as follows:

$$\frac{\partial J}{\partial w_i} = \sum_{k=1}^{d} \frac{\partial J}{\partial w_{i,k}},$$

In which d implies a dimension of word i.

If w_i is a vector for (x_1, x_2, x_3), $\frac{\partial J}{\partial w_i}$ would be:

$$\left(\frac{\partial J}{\partial w_{i,1}} = (1,0,0)\right) + \left(\frac{\partial J}{\partial w_{i,2}} = (0,1,0)\right) + \left(\frac{\partial J}{\partial w_{i,3}} = (0,0,1)\right) \quad (8)$$

Therefore, a vector of $(1,1,1)$. Hence, the manner of calculating derivatives is based on a word wherein GD is employed by learning rate alpha $(= 0.5)$, to train $word2Vec$ module. Consequently, when the module is trained, the words with similar meanings are extracted with producing an arbitrary word as input. In businesses, words with similar meanings such as business, Market, industry, products, share market, stock, etc.

C. Sarcasm Detection using CNN-RNN Technique

The extracted feature word vectors are fed into the CNN-RNN technique for the classification of sarcasm. RNN is a kind of NN which preserves internal hidden state for modelling dynamic temporal behaviour of series using random lengths by directed cyclic relations among its unit. It could be taken into account as a hidden Markov method extension which applies nonlinear transition function and can able to model long-term temporal dependency. LSTM prolongs RNN by including a forget gate f for controlling either to forget the present state; an input gate i for indicating whether it read the input; an output gate o for controlling either to output the states [23]. That gate enables LSTM for learning long-term dependencies in a series, and also facilitates it for optimizing since that gate helps the input signal to efficiently broadcast via the recurrent hidden state $r(t)$ without influencing the output. Also, LSTM efficiently handles the gradient exploding or vanishing problems which usually appear in RNN training. Fig. 2 illustrates the framework of CNN model.

$$\begin{aligned} x_t &= \delta\left(U_r . r(t-1) + U_w w_k(t)\right) \\ i_t &= \delta\left(U_{i_r} r(t-1) + U_{i_\omega} w_k(t)\right) \\ f_t &= \delta\left(U_{f_r} r(t-1) + U_{f_\omega} w_k(t)\right) \\ o_t &= \delta\left(U_{o_r} r(t-1) + U_{o_w} w_k(t)\right) \\ r(t) &= f_t \odot r(t-1) + i_t \odot x_t \\ o(t) &= r(t) \odot o(t) \end{aligned} \quad (9)$$

In which $\delta(.)$ denotes an activation function, $\odot$ indicates the product using gate value, and different W matrices are learned parameters. They use the rectified linear unit (ReLU) as the activation function in this performance. A new CNN-RNN architecture is employed for multi label classification problems. It consists of: The CNN extract semantic representation from the image; the RNN models label or image relations and label dependency. The recurrent, label and image depictions are proposed to the similar low dimension space for modelling the label redundancy and the image text relation.

Fig. 3 demonstrates the structure of RNN model. The RNN method is applied as a compact but a strong representation of the label co-existence dependencies in this space. It takes the embedding of the predictive labels at every time step and maintains a hidden state for modelling the label's co-existence data. The a priori likelihood of a label provided the previous prediction label could be calculated based on their dot product with the addition of recurrent and image embeddings.

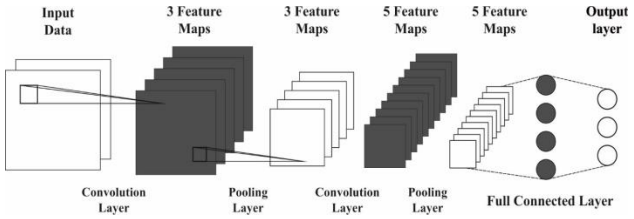


Fig. 2. Structure of CNN.

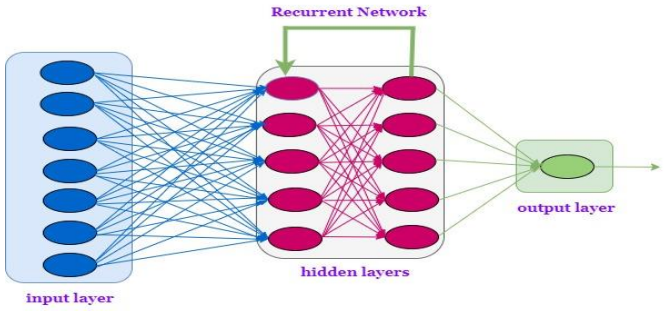


Fig. 3. Architecture of RNN.

A label k is denoted as a one-hot vector $e_k = [0, \dots 0, 1, 0, \dots, 0]$, i.e., 1 at kth position, and 0 somewhere else. The label embedding could be attained by multiplying the one-hot vector using a *label embedding matrix* U_l. The kth row of U_ι denotes the label embedding of label k.

$$w_k = U_l . e_k. \tag{10}$$

The dimension of w_k is generally lower compared to the amount of labels. The recurrent layer takes the label embedding of the previous prediction label, and models the co-existence dependency in its hidden *recurrent state* by learning nonlinear function:

$$o(t) = h_o\big(r(t-1), w_k(t)\big), r(t) = h_r\big(r(t-1), w_k(t)\big) \tag{11}$$

whereas $r(t)$ and $o(t)$ represents the hidden state and output of the recurrent layer at the time step t, correspondingly, $w_k(t)$ indicates the label embedding of t-th label in the predictive path, and $h_o(.)$, h_r $(.)$ signifies the nonlinear RNN function. The image depiction and output of recurrent layer are proposed to the similar low dimension space as the label embedding.

$$x_t = h(U_o^x o(t) + U_I^x I), \tag{12}$$

whereas U_o^x & U_I^x denotes the prediction matrix for image depiction and recurrent layer output, correspondingly. The column count of U_o^x & U_I^x indicates the similar as label embedding matrix $U_l . I$ represents the CNN image depiction. They would display in second that the learned joint embedding efficiently characterizes the significance of labels and images. Lastly, the label score could be calculated by multiplying the transpose of U_l & x_t for computing the distances among x_t and every label embedding.

$$s(t) = U_l^T x_t. \tag{13}$$

The prediction label likelihood could be calculated by soft max regularization on the scores. Fig. 4 depicts the architecture of LSTM model.

D. Hyperparameter Optimization using TLBO Algorithm

At the final stage, the learning rate of the CNN-RNN technique is optimally chosen by the use of TLBO algorithm in such a way that the sarcasm detection outcome gets increased. TLBO technique is a novel type of metaheuristic approach which is dependent upon teaching– learning model. It can be established by Rao et al. [24] for solving optimization issues. It can be simulated as feeding the knowledge in a class where students initially gain information from teacher and next with mutual interface. The TLBO technique has population based optimized technique where the set or class of students regarded as population. Therefore, the student of class signifies the possible solution of difficulty. The TLBO technique includes two stages as given below.

1) Teacher level: This level defines the learning of student from teacher. The teachers attempt for improving the knowledge level of student and uses for obtaining optimum marks. However, the student gains information and attain marks based on quality of teaching distributed as teacher and quality of student existing in the class. In order to simulate, supposing there are 'n' amount of subjects ($j = 1,2, \dots, n$) existing to 'N_p' amount of students (population size, $i = 1,2, \dots N_p$). In some teaching-learning cycles (iteration, $k = 0,1,2, \dots I_n$), M_j^k represents the mean outcome of students in specific subject 'j'. The teacher is one of the skilled, experienced, and extremely learned person in society. For simulating this model, an optimum student (possible solution) in total population was regarded as teacher. The variance among the outcome of teacher and the mean outcome of students in subject 'j' is provided as:

$$D_j^k = r\big(X_{T,j}^k - T_F M_j^k\big) \tag{14}$$

where T_F implies the teaching influence that decided the value of mean that altered and r implies the arbitrary number in range 0 to 1. T_F signifies not parameter of TLBO technique and their value is either be one or two [25]. The possible solution (student) is enhanced by moving its places near the place of an optimum possible solution (teacher) by taking into account the present mean value of possible solution. For simulating this detail, the i^{th} possible solution in the population at k^{th} teaching-learning cycle is upgraded based on subsequent written as:

$$X_{new,i,j}^k = X_{old,i,j}^k + D_j^k \tag{15}$$

When $X_{new,i}^k$ implies the superior to $X_{old,i}^k$, then $X_{new,i}^k$ is recognized; Then it can be rejected. Every accepted possible solution is continued and these developed the input to student phase.

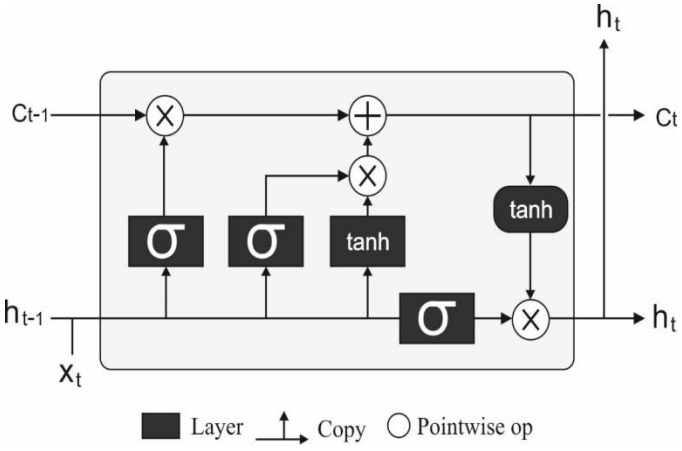


Fig. 4. Framework of LSTM.

2) Student level: Here, the student gains information with mutual communication. The students interrelate arbitrarily with another student of class for improving knowledge. Therefore, when the student (v) has superior to student (u), afterward student (u) is stimulated near student (v). Then, student (u) was stimulated away from student (v). The learning viewpoint of this phase is provided here. Two students (possible solution, X_u^k, X_v^k) are arbitrarily elected from class (population), where u, v are 2 integers arbitrary number go to [1, N_p] and $u \neq v$.

If $F(X_u^k) > F(X_v^k)$

$$X_{new\_SP,u,j}^k = X_{u,j}^k + r(X_{u,j}^k - X_{v,j}^k)$$

Else

$$X_{new\_SP,u,j}^k = X_{u,j}^k + r(X_{v,j}^k - X_{u,j}^k)$$

Endif

where F(x) implies the fitness function (FF) which is utilized for finding the fitness value of possible solutions, $X_{new\_SP,u,j}^k$ refers the jth design variable of altered possible solutions from student level at kth teaching-learning iteration.

Afterward, the fitness value of $X_{new\_SP,u}^k$ is estimated.

If $F(X_{new_{SP},u}^k) > F(X_{new,u}^k)$

$$X_{new,u}^k = X_{new_{SP},u}^k$$

Else

$$X_{new,u}^k = X_{new,u}^k$$

End if

IV. Performance Validation

A. Implementation Setup

The DLE-SDC technique is implemented on Python 3.6.5 tool with additional packages such as tensorflow-gpu==2.2.0, pandas, nltk, tqdm, scikit-learn, matplotlib, seaborn, pretty table, pyqt5==5.14, pycm, and numpy==1.19.5. The DLE-SDC technique is tested using a News Headlines Dataset For Sarcasm Detection dataset from Kaggle Repository [26]. The execution process of the DLE-SDC technique is given in Appendix (Fig. 13 to 16).

B. Results Analysis

This section examines the sarcasm detection performance of the DLE-SDC technique against several aspects. The DLE-SDC technique is investigated interms of different measures namely precision, recall, accuracy, and F-measures. The confusion matrix generated by the DLE-SDC technique on the classification of sarcasm is depicted in Fig. 5. The figure showcased that the DLE-SDC technique has classified a total of 14166 instances into non-sarcastic and 12639 instances into sarcastic ones.

A brief classification results analysis of the DLE-SDC technique with other DL techniques takes place in Table I and Fig. 6. From the resultant values, it is noticeable that the CNN-TLBO algorithm has offered lower classification outcomes with the precision of 0.9021, recall of 0.8946, accuracy of 0.8974, and F-measure of 0.8943. Then, the RNN-TLBO algorithm has gained somewhat increased performance with the precision of 0.9135, recall of 0.9036, accuracy of 0.9067, and F-measure of 0.9035. Eventually, the CNN-RNN model has obtained moderately closer outcome with the precision of 0.9275, recall of 0.9160, accuracy of 0.9192, and F-measure of 0.9182. However, the DLE-SDC technique has outperformed all the other DL models with the precision of 0.9406, recall of 0.9401, accuracy of 0.9405, and F-measure of 0.9403.

Accuracy analysis of the DLE-SDC technique is investigated under varying numbers of epochs in Fig. 7. The figure showcased that the training and training accuracy values get increased with an increase in epoch count. Particularly, validation accuracy is found to be superior to training accuracy.

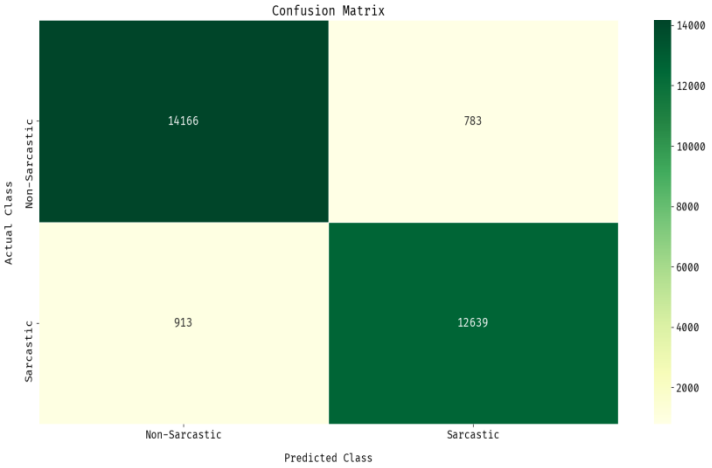


Fig. 5. Confusion Matrix of the DLE-SDC Model.

TABLE I. Results Analysis of Proposed DLE-SDC Model with Various Deep Learning Models

Methods	Precision	Recall	Accuracy	F-Measure
Proposed DLE-SDC	0.9406	0.9401	0.9405	0.9403
CNN-RNN	0.9275	0.9160	0.9192	0.9182
RNN-TLBO	0.9135	0.9036	0.9067	0.9035
CNN-TLBO	0.9021	0.8946	0.8974	0.8943

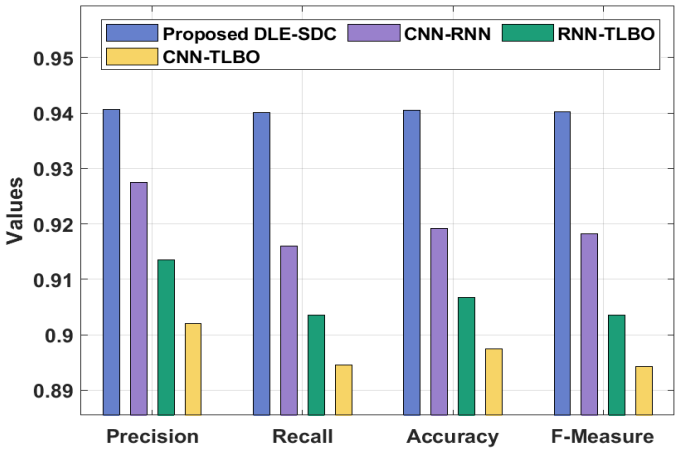


Fig. 6. Result Analysis of DLE-SDC Model with Different Measures.

Fig. 7. Accuracy Graph on Proposed DLE-SDC Model.

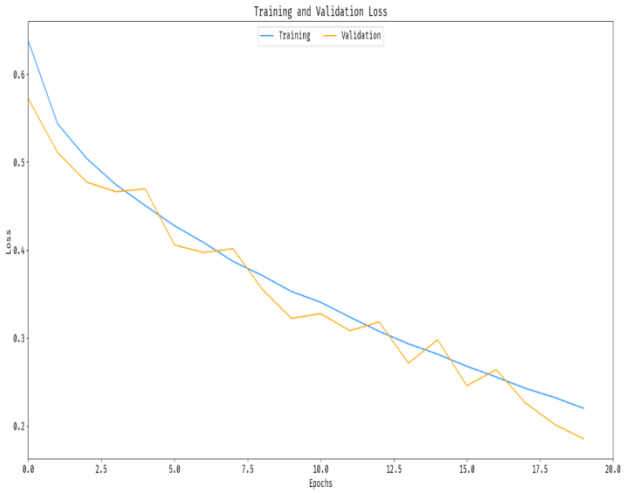


Fig. 8. Loss Graph on Proposed DLE-SDC Model.

A loss graph analysis of the DLE-SDC technique is examined under variable number of epochs in Fig. 8. The figure has shown that the training and training loss values get reduced with a rise in number of epoch. Particularly, the validation loss seems to be lower than the training loss of the DLE-SDC technique.

Fig. 9 examines the ROC analysis of the proposed DLE-SDC technique on the applied dataset. The figure portrayed that the proposed DLE-SDC technique has accomplished better performance with a maximum ROC of 0.98.

Finally, a comprehensive comparative study of the DLE-SDC technique with other techniques takes place in Table II [27]. Fig. 10 displays the precision analysis of the DLE-SDC technique with other techniques. The figure portrayed that the GRNN and VLSTM techniques have offered ineffective outcomes with the least precision of 0.663 and 0.673 respectively. Also, the E-BiLSTMA and ALSTM techniques have showcased slightly improved outcomes with the precision of 0.684 and 0.687 respectively. Moreover, the VCNN, NBOW, E-BiLSTM, and E-BiLSTMMF techniques have reached a moderately closer precision of 0.71, 0.712, 0.759, and 0.778 respectively. Furthermore, the MMNSS and A2Text-Net techniques have obtained competitive outcomes with the precision of 0.857 and 0.917. However, the proposed DLE-SDC technique has resulted in superior outcomes with a maximum precision of 0.941.

Fig. 11 showcases the recall analysis of the DLE-SDC approach with other methods. The figure demonstrated that the NBOW and GRNN approaches have offered ineffective results with the minimum recall of 0.623 and 0.647 correspondingly. In line with this, the VCNN and VLSTM manners have exhibited somewhat increased results with the recall of 0.671 and 0.672 correspondingly. Furthermore, the ALSTM, E-BiLSTMA, E-BiLSTMF, and E-BiLSTM methodologies have reached a moderately closer recall of 0.686, 0.708, 0.735, and 0.750 correspondingly. Along with that, the MMNSS and A2Text-Net methods have attained competitive results with the recall of 0.892 and 0.910. Eventually, the projected DLE-SDC approach has resulted in higher results with the maximal recall of 0.940.

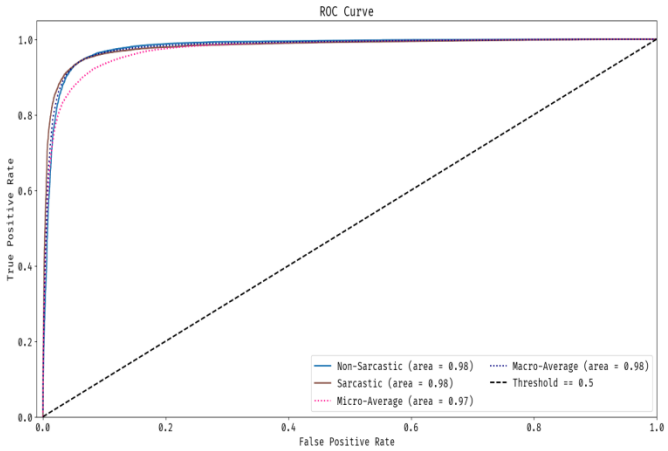


Fig. 9. ROC Analysis on Proposed DLE-SDC Model.

TABLE II. RESULTS ANALYSIS OF EXISTING WITH PROPOSED DLE-SDC MODEL IN TERMS OF DIFFERENT MEASURES

Methods	Precision	Recall	F-Measure
Proposed DLE-SDC	0.941	0.940	0.940
MMNSS	0.857	0.892	0.871
NBOW	0.712	0.623	0.641
VCNN	0.710	0.671	0.685
VLSTM	0.673	0.672	0.672
ALSTM	0.687	0.686	0.687
GRNN	0.663	0.647	0.654
E-BiLSTM	0.759	0.750	0.759
E-BiLSTMF	0.778	0.735	0.753
E-BiLSTMA	0.684	0.708	0.694
A2Text-Net	0.917	0.910	0.900

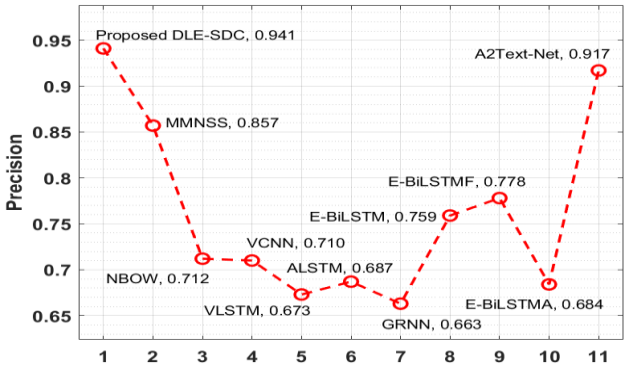


Fig. 10. Precision Analysis of DLE-SDC Model with Existing Techniques.

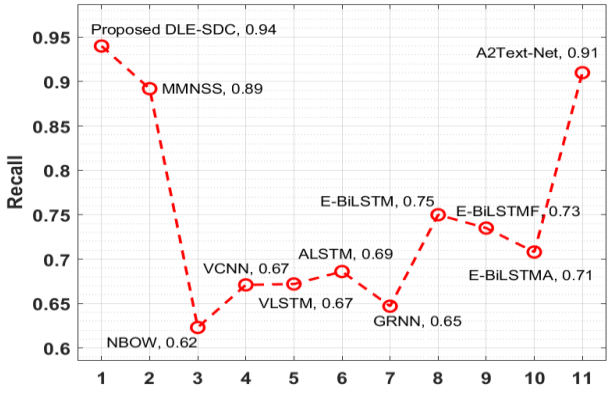


Fig. 11. Recall Analysis of DLE-SDC Model with Existing Techniques.

Fig. 12 depicts the F-measure analysis of the DLE-SDC method with other algorithms. The figure outperformed that the NBOW and GRNN methods have offered ineffective outcomes with the worse F-measure of 0.641 and 0.654 correspondingly. Similarly, the VLSTM and VCNN techniques have outperformed somewhat increased results with the F-measure of 0.672 and 0.685 correspondingly. In line with, the ALTSM, E-BiLSTMA, E-BiLSTMMF, and E-BiLSTM algorithms have reached a moderately closer F-measure of 0.687, 0.694, 0.753, and 0.759 respectively. Furthermore, the MMNSS and A2Text-Net methodologies have obtained competitive outcomes with the F-measure of 0.871 and 0.900. However, the proposed DLE-SDC approach has resulted in maximum results with the superior F-measure of 0.940.

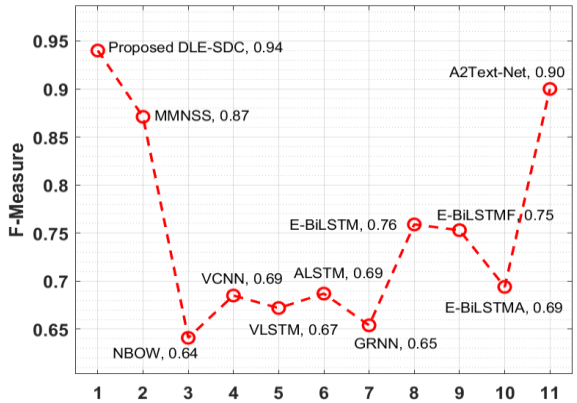


Fig. 12. F-measure Analysis of DLE-SDC Model with Existing Techniques.

From the above mentioned results analysis, it is observed that the DLE-SDC technique has accomplished maximum sarcasm detection performance and can be employed to detect sarcasm in online social media content.

V. Conclusion

This paper has presented a new DLE-SDC technique to identify and classify the sarcasm using DL technique. The proposed DLE-SDC technique comprises different stages of operations such as pre-processing, word vector representation, CNN+RNN based classification, and TLBO based hyper parameter optimization. Besides, the CNN-RNN technique involves the BiLSTM model of the detection and classification of sarcasm. In order to increase the sarcasm detection performance of the CNN-RNN model, TLBO algorithm is applied to determine the optimal learning rate of the presented CNN-RNN model and it is mainly used to boost the detection performance to a maximum extent. A wide range of simulations take place on benchmark datasets and validate the results interms of different measures. The simulation outcomes pointed out the supremacy of the DLE-SDC technique over the recent state of art techniques. As a part of future work, the sarcasm detection performance can be extended to the design of feature selection and clustering techniques.

References

[1] N. Majumder, S. Poria, H. Peng et al., "Sentiment and sarcasm classification with multitask learning," IEEE Intelligent Systems, vol. 34, no. 3, pp. 38–43, 2019.

[2] G. Deliens, K. Antoniou, E. Clin, and M. Kissine, "Perspective-taking and frugal strategies: evidence from sarcasm detection," Journal of Pragmatics, vol. 119, pp. 33–45, 2017.

[3] A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. AbdelBasset, "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network," IEEE Access, vol. 7, pp. 23319–23328, 2019.

[4] S. K. Bharti and S. B. Korra, "Sarcasm detection in twitter data: a supervised approach," in Semantic Web Science and RealWorld applications, M. D. Lytras, N. Aljohani, E. Damiani, and K. T. Chui, Eds., pp. 246–272, IGI Gobal, Hershey, PA, USA, 2019.

[5] D. Ghosh, W. Guo, and S. Muresan, "Sarcastic or not: word embeddings to predict the literal or sarcastic meaning of words," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1003–1012, Lisbon, Portugal, 2015.

[6] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4619–4629, Florence, Italy, 2019.

[7] Kreuz, R.J.; Caucci, G.M. Lexical influences on the perception of sarcasm. In Proceedings of the Workshop on Computational Approaches to Figurative Language, Association for Computational Linguistics, Rochester, NY, USA, 26 April 2007; pp. 1–4.

[8] Joshi, A.; Sharma, V.; Bhattacharyya, P. Harnessing context incongruity for sarcasm detection. In Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, Beijing, China, 26–31 July 2015; pp. 757–762.

[9] Ghosh, A.; Veale, T. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In Proceedings of the 2017 Conference on EMNLP, Copenhagen, Denmark, 7–11 September 2017; pp. 482–491.

[10] Ghosh, D.; Fabbri, A.R.; Muresan, S. Sarcasm analysis using conversation context. Comput. Linguist. 2018, 44, 755–792.

[11] Nayel, H., Amer, E., Allam, A. and Abdallah, H., 2021, April. Machine learning-based model for sentiment and sarcasm detection. In

Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 386-389).

[12] Kumar, H.K. and Harish, B.S., 2018. Sarcasm classification: a novel approach by using content based feature selection method. Procedia computer science, 143, pp.378-386.

[13] Chatterjee, N., Aggarwal, T. and Maheshwari, R., 2020. Sarcasm detection using deep learning-based techniques. In Deep Learning-Based Approaches for Sentiment Analysis (pp. 237-258). Springer, Singapore.

[14] Razali, M.S., Halin, A.A., Ye, L., Doraisamy, S. and Norowi, N.M., 2021. Sarcasm Detection Using Deep Learning With Contextual Features. IEEE Access, 9, pp.68609-68618.

[15] Rajeswari, K. and ShanthiBala, P., 2018. Sarcasm detection using machine learning techniques. Int J Recent Sci Res., 9, pp.26368-26372.

[16] Zhang, M., Zhang, Y. and Fu, G., 2016, December. Tweet sarcasm detection using deep neural network. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers (pp. 2449-2460).

[17] Akula, R. and Garibay, I., 2021. Interpretable Multi-Head Self-Attention Architecture for Sarcasm Detection in Social Media. Entropy, 23(4), p.394.

[18] Kastrati, Z., Ahmedi, L., Kurti, A., Kadriu, F., Murtezaj, D. and Gashi, F., 2021. A Deep Learning Sentiment Analyser for Social Media Comments in Low-Resource Languages. Electronics, 10(10), p.1133.

[19] Das, S. and Kolya, A.K., 2021. Parallel Deep Learning-Driven Sarcasm Detection from Pop Culture Text and English Humor Literature. In Proceedings of Research and Applications in Artificial Intelligence (pp. 63-73). Springer, Singapore.

[20] Sundararajan, K. and Palanisamy, A., 2020. Multi-rule based ensemble feature selection model for sarcasm type detection in twitter. Computational intelligence and neuroscience, 2020.

[21] Kumar, A., Dikshit, S. and Albuquerque, V.H.C., 2021. Explainable Artificial Intelligence for Sarcasm Detection in Dialogues. Wireless Communications and Mobile Computing, 2021.

[22] Liu, Liyuan, Jennifer Lewis Priestley, Yiyun Zhou, Herman E. Ray, and Meng Han. "A2text-net: A novel deep neural network for sarcasm detection." In 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI), pp. 118-126. IEEE, 2019.

[23] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. and Xu, W., 2016. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2285-2294).

[24] Rao RV, Savsani VJ, Vakharia DP. Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. Computer-Aided Des 2011;43(3):303–15.

[25] Gill, H.S., Khehra, B.S., Singh, A. and Kaur, L., 2019. Teaching-learning-based optimization algorithm to minimize cross entropy for Selecting multilevel threshold values. Egyptian Informatics Journal, 20(1), pp.11-25.

[26] https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection

[27] Ren, L., Xu, B., Lin, H., Liu, X. and Yang, L., 2020. Sarcasm detection with sentiment semantics enhanced multi-level memory network. Neurocomputing, 401, pp.320-326.

APPENDIX

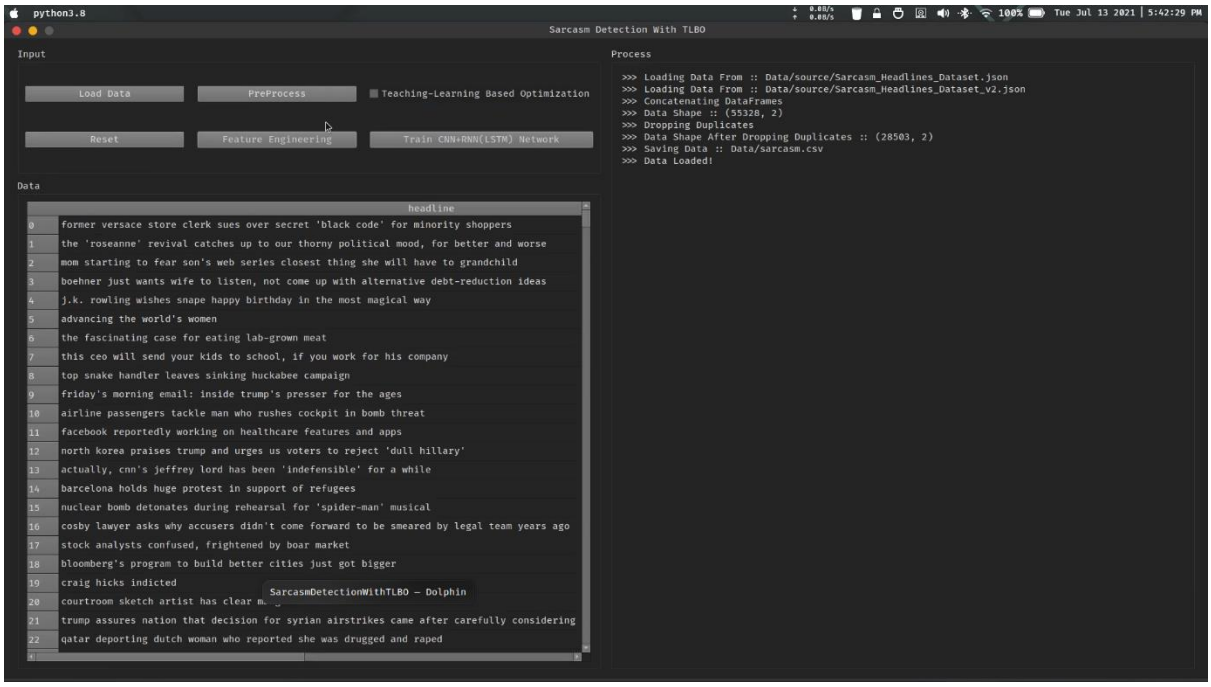


Fig. 13. Loading Dataset Module

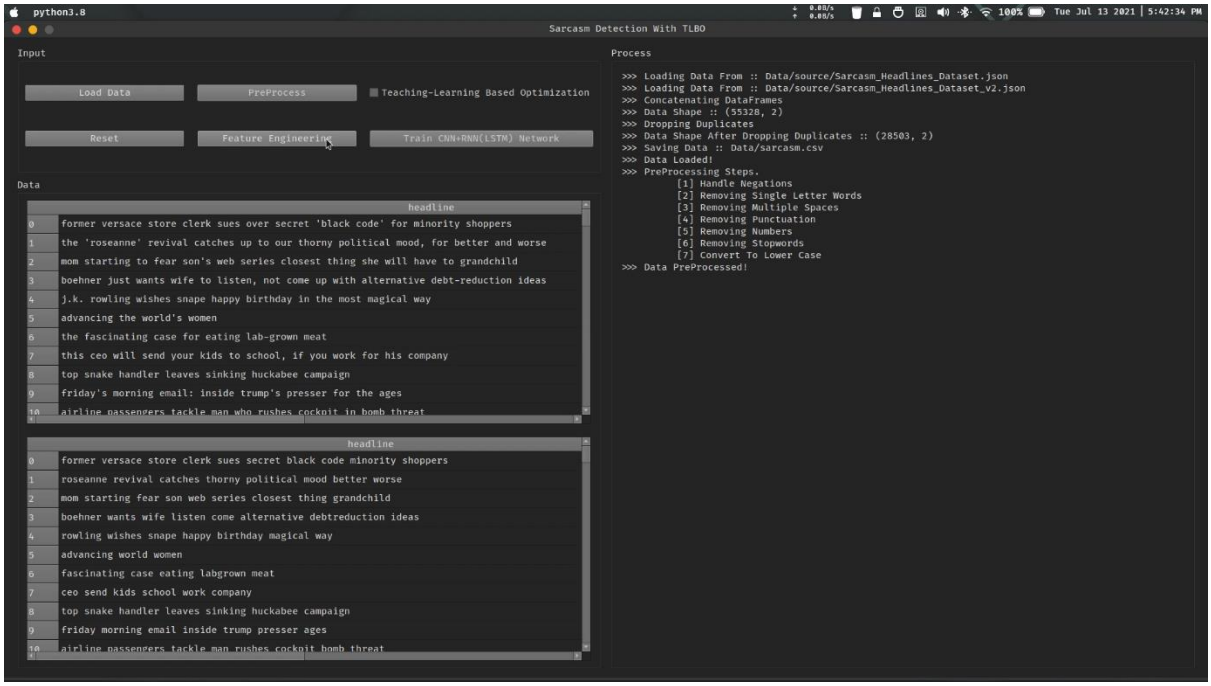


Fig. 14. Preprocessing Module

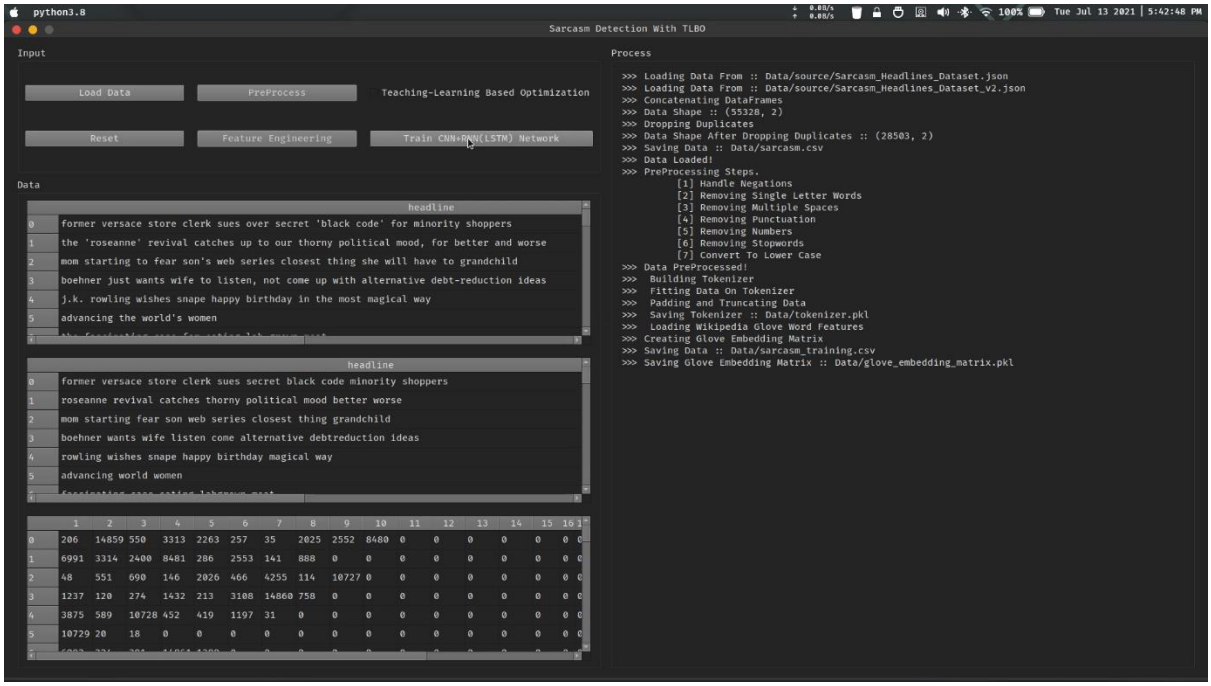


Fig. 15. Feature Extraction Module

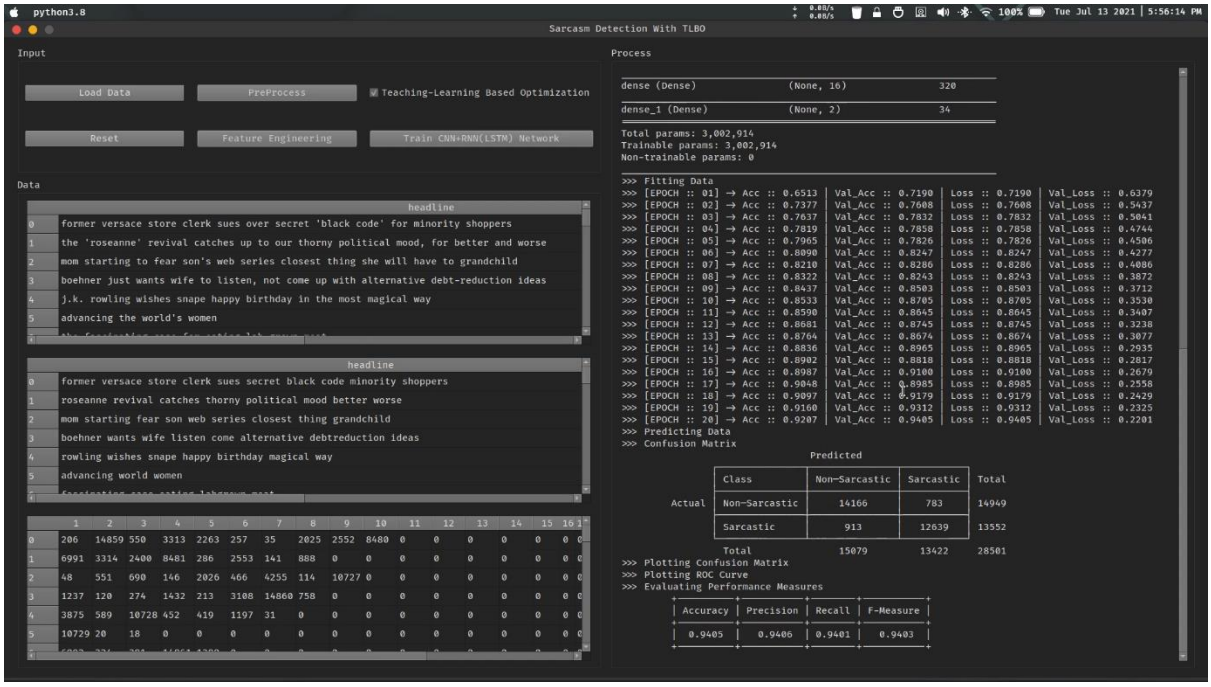


Fig. 16. Classification Module

DBTechVoc: A POS-tagged Vocabulary of Tokens and Lemmata of the Database Technical Domain

Jatinderkumar R. Saini[1]*, Ketan Kotecha[2], Hema Gaikwad[3]
Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, India[1, 3]
Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International Deemed University, Pune, India[2]

***Abstract*—Vocabulary of a language has a great role to play in the Natural Language Processing (NLP) applications. Such applications make use of lists like stop-word list, general service list, academic word list and technical domain word list. The technical domain word list differs with each domain and though it is available for fields like medicine, biology, computer science, physics and law, the domain of databases in specific has still not been explored. For the first time, we propose technical vocabulary comprising of POS-tagged unigram tokens and POS-tagged unigram lemmata for the technical domain of databases. This vocabulary has been called DBTechVoc with a coined term. Notably, the multi-word phrases have also been considered, without their further tokenization, to maintain their semantics. The empirical results, with more than 1000 high quality research papers collected over a period of 45 years from 1976 to 2021, prove that the technical general word list of the domain of computer science is different from the technical and specific word list of the domain of databases. The overlap was found to be less than 2%. The research titles use 6% Rainbow stop words while 13% of the words used for the research paper titles are inflectional forms of lemmata.**

***Keywords*—Database; lemma; part-of-speech (POS); technical word list; token; unigram; vocabulary**

I. Introduction

It has been empirically proved by Liu and Nation [1] that in order to comprehend a piece of text, at least 95% of the words should be recognized by the reader. In fact, this concept could be applied equally well to the listeners of a natural language too. Any person's knowledge of a language is just limited by the knowledge of the vocabulary of that language. It is needless to mention that though grammar of a language has an important role to play too, it is the number of words known to a reader or listener that contributes to the comprehension of the semantics of a language. There are a number of specific terms like tokens, lemmata and stop words, just to name a few, which are used by the linguists, computational linguists as well as those working in the area of Natural Language Processing (NLP).

The importance of title of a research paper cannot be undermined. Several research works like those of Dewan and Gupta [2], Tullu [3], Mack [4] and Karagel and Karagel [5] have advocated and elaborated the importance of title of the research paper as a gist of the paper contents. Soler [6] conducted an exploratory study dedicated to the titles of scientific research papers. Hengl and Gould [7] emphatically highlighted that the title of the research papers should tend to clearly indicate the main contents of the research paper in addition to the actual discoveries discussed in the paper.

Any text used to convey the necessary semantics consists of words. The process used for separating the individual units of this text is called Tokenization and the units so received are called Tokens [8]. These tokens may in turn be formed of single word, two words, etc. which are technically referred to as unigrams, bigrams, etc. respectively. Unlike Kyle [9], we have not considered the relevance of single, double, etc. unigrams. Also, the consideration of the unigrams in the present research work is with respect to the number of words in a sentence rather than the number of letters in a word. More specifically, for a period of 45 years from 1976 to 2021, we have considered 1031 titles of the database domain related research papers as the sentences and extracted unigrams as well as lemmata from these titles. The process of lemmatization is deployed to find the base morphological form of a word [10]. This form is called 'lemma' if it is singular, and 'lemmas' or 'lemmata' if it is plural. Similarly, the Part-Of-Speech (POS) tagging is done in order to group the various tokens into different categories as well as to provide more information on the role of such tokens when used as words in a sentence [11]. The present research work makes use of Lemmatizer [12] provided by the Stanford University and the POS-tagger [13] provided by the University of Copenhagen. The POS-tagger [14] provided by the Princeton University has also been used.

Smith [15] has discussed three main types of lists, viz. Academic, General and Technical. He defines the Academic Vocabulary is the list containing words which could be used for discourse in the academic world including the usage during conferences. He defines the General Vocabulary as consisting of the most frequently used words for a language. Similarly, he advocates that the Technical Vocabulary consists of the discipline-specific words. In wake of this context, the present research work deals with proposing the lists which fit in the category of Academic Vocabulary and Technical Vocabulary. It does not fit in the definition of General Vocabulary as we have not considered the words based on their frequency. We have presented the token list as well as the lemmata list, both of which are POS-tagged, towards the academic and technical categories of words for the technical database domain.

Rest of the paper is structured as follows: Section 2 presents the pertinent literature review. Section 3 elaborates the methodology, followed by section 4 presenting the results and discussion. The paper ends with the last Section 5 on conclusion and limitations of the present research work. Many

*Corresponding Author.

application areas and directions of future work are also presented in the last section.

II. Literature Review

Ever since the researchers recognized the importance of the vocabulary of a language, they have been working for the generation and in the field of word-lists. The research has gained more interest in the wake of several developments including the growth of various interdisciplinary fields like Computational Linguistics (CL), Natural Language Processing (NLP) and Foreign Language Understanding (FLU), among others. First of its kind, a general service list, comprising of most commonly used English words was proposed way back in 1953 by West [16]. This list has seen two updates, viz. new-GSL by Brezina and Gablasova [17] and NGSL by Browne et al. [18], in recent times with inclusion of additional words owing to consideration of bigger corpora respectively for more than 12 billion words and 2 billion words respectively. It is important to note that where GSL by West [16] contained some 2000 words, NGSL by Browne et al. [18] contained 2818 lemma words. Gilner and Morales [19] used the existing GSL and presented a speech-based analysis of the words in the list. Gilner [20] also presented an introductory note on the description of GSL with an aim to aid and ease its comprehension. Nation and Waring [21] argued that the most part of the text is actually composed of only a few words which occur frequently in the text.

The stop-word lists of the various natural languages contribute to the creation of a language itself. Though such stop words or noise words are believed to statistically irrelevant and mostly useless from point of view of NLP applications too, they enable the spoken use and representation of the vocabulary of a language through speech, dialects and scripts. They help in putting the vocabulary words together to make sense to the listener and reader. This way they contribute to a special vocabulary domain in its own right. Researchers have presented various types of stop-word lists as well as those for several languages. Researchers have also worked a lot on the analysis and classification of stop-word lists for various languages. Fayaza and Farhath [36] presented a stop word list for Tamil language. Similarly, Kaur and Saini [22, 23] worked for the stop-word list of Punjabi language, Rakholia and Saini [24, 25] worked for the stop-word list of Gujarati language while Raulji and Saini [31, 32] worked for the stop-word list of Sanskrit language. A stop word list based on the Rainbow statistical text has also been presented by Shuson [38].

Similar to the concept of GSL and NGSL, Coxhead [27] proposed the concept of an Academic Word List (AWL). However, Hancioglu et al. [26] argued that it is inappropriate to treat AWL and GSL as separate lists. Billurog˘lu and Neufeld [28] used a rather simplistic approach of list generation by filtering out the unique common words from the corpus created by the merging of all existing and commonly used lists. The concept of various lists has gained importance and interest as the various words contained in such lists provide a glimpse into the vocabulary of a language or a specific domain thereof. It is the knowledge of this vocabulary and understanding of words which helps one to understand and learn a language.

In addition to the core research works on various types of lists, researchers have also explored other similar and pertinent domains. For instance, a number of methods exist for the extraction of the terms from the scripted version of the language. The extracted tokens are in turn used by various downstream operations in the field of CL and NLP. Bakaric et al. [29] evaluated many such methods for the German language. Choy [30] proposed an innovative method for generation of stop word list by making use of combinatorial values. Venugopal et al. [33] presented lemmata for the Hindi corpus stop words while Saini and Rakholia [34] presented a detailed statistical analysis for such lists for various international languages.

Saed et al. [39] presented a lemmata list of the various categories related to biological and medical sciences including for the classes of diseases and the recent COVID-19 outbreak. Das et al. [40] used various sources and presented the technique of generating a list of words for the specific domain of Finance. They presented a typical comparison and contrast of their lexicographic approach with the conventional machine learning based approaches. Ahsanuddin et al. [41] attempted to create a list of words for the vocabulary learning by the students aiming to learn languages like Indonesian, English, German and Arabic. They used nearly 380 Thousand tokens for the corpus creation. Joensuu [42] presented an innovative description of the lists of menus and recipes for the culinary domain. The language researched by the author was Finnish.

Using the lists like AWL by Coxhead [27], Wingrove [43] attempted to analyze the introduction of TED talks for English learners. The list of words extracted from the talks and other lectures was analyzed for the possibility of vocabulary enrichment of the language learners. On the sidelines, he also analyzed the richness of such talks from the perspective of the usage of different lexicons. Alasmary [44] presented a technical list of words for the domain of mathematics. He sourced the corpus from the textbooks of the mathematics course at the graduate-level of students.

Though a list of the database terms is provided by raima.com [35], it consists of only a limited 150 terms, without POS and more in the form of a dictionary. Also it has not made use of lemmatization to present the lemmata list. The present research work considers all such points by providing an improved set of lists. Smith [15] has presented a few subject-specific lists like for Medicine, Law, Computer Science, Physics, Chemistry and Accounting but he has not presented a specific technical word list for the subject of Database which happens to be a sub-field under the umbrella of Computer Science. Also, the Computer Science subject related word list provided by him is very different from the vocabulary used in the Database domain.

After a thorough literature review, it was concluded that though several types of lists like stop-word lists of different types and for different languages, general service lists of various types and many academic word lists exist, the area of technical domain word lists is rather unexplored. This is particularly true for the highly technical domains like that of databases. Additionally, as no such list exists for the specific field of databases, there is no research work which has

elaborately annotated such a list with Parts-of-Speech (POS). In order to bridge this gap, this research work presents a technical domain word list for the domain of databases. It is remarkable that as the field of databases itself is a sub-set of the field of computers, the proposed lists could also be used with backward inclusion in the technical domain list of the parent field of computers in general. Hence, the contributions of the present research work are manifold in terms of presentation of vocabulary and word lists. In the increasing order of generality, firstly, it presents a list of vocabulary words for databases, secondly it presents a technical word list and finally it also presents the vocabulary word list for the field of computer science and engineering as well as information technology.

III. Methodology

All the executions of the multiple codes needed at different junctures of the present research work were done using the open source Java programming language with version 17.0.1 2021-10-19 LTS for Java Development Kit (JDK), build 17.0.1+12-LTS-39 for the Standard Edition (SE) Runtime Environment and build 17.0.1+12-LTS-39 with mixed mode and sharing features for Java HotSpot(TM) 64-bit server Virtual Machine (VM). The execution was done on a machine with Intel(R) Core(TM) i3-8145U CPU with 2.10 GHz, 8 GB RAM and a licensed Windows 10 Pro 64-bit operating system.

In order to create a subject-specific vocabulary of the technical domain of databases, two Part-Of-Speech (POS) tagged lists were created. The diagrammatic representation of the process is depicted in Fig. 1. As a first step, the list of titles of research papers published in the field of databases from 1976 to 2021 were collected. In order to assure the duration of publications, quality of research publications and the scope of the present research work, only the papers published in the ACM Transactions on Database Systems (ACM TODS) [37] were considered. The collected list of 1031 titles from all the research papers of this duration was subjected to tokenization in order to extract the words from the titles. The tokenization was performed without considering the case of the words but maintaining the Multi-word phrases (MWP). Only unigrams were considered for the present research work. The resultant list consisted of 8139 words. This list could be considered a technical word list.

Cleaning was performed in this list to remove various noise words in context of the present research work. This constituted removal of unigrams like years (e.g. 1977, 2005, etc.), numbers (e.g. 3, 6, etc.) and special characters (e.g. *, #, etc.). The resultant list with 7994 words was used to find unique tokens. It is noteworthy that this point onwards, in order to emphasize the unique words in the list, we term the words as tokens. The count of such tokens was 1900. Stop words were removed from this list. We considered the 526 stop words provided by the standard Rainbow Stop Word List [38] for the present research work. The Rainbow Stop Word List had no MWP and its snapshot is provided in Table I. The resultant list was the refined technical list containing unique, lower-cased and non-stop-word 1791 tokens.

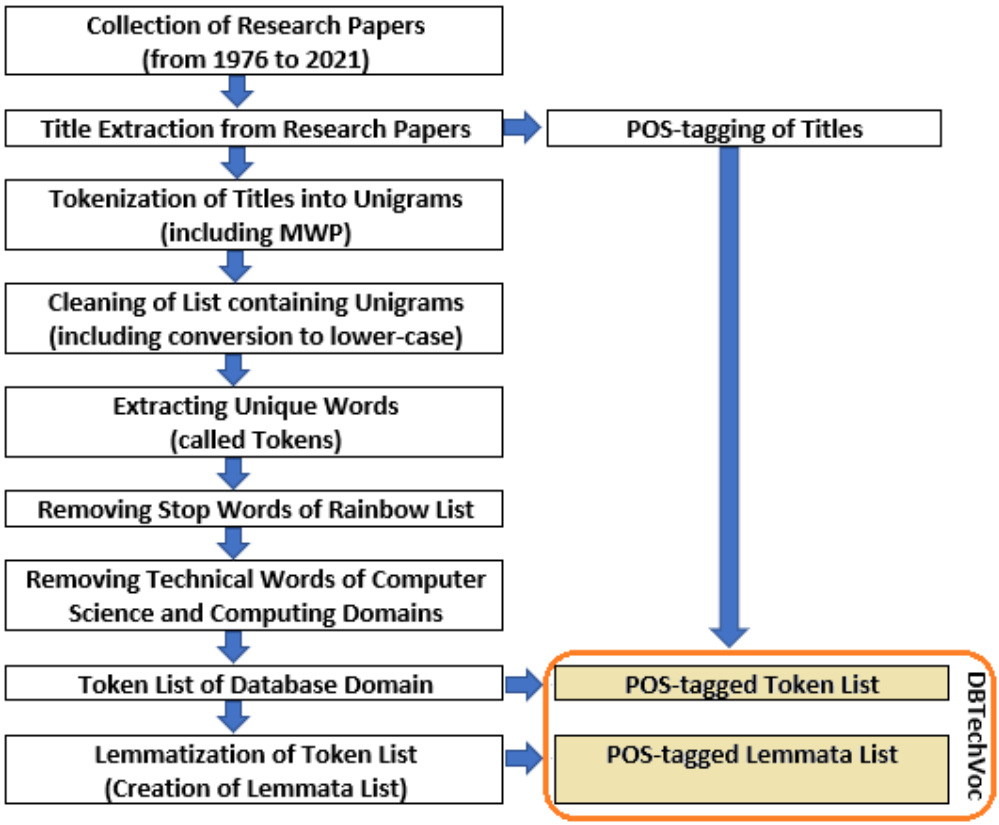


Fig. 1. Diagrammatic Representation of the Methodology to Create DBTechVoc.

The aim of present research work was the development of a database-specific vocabulary. Hence, at this stage, we referred another standard technical word list [15] containing the words from general domains of computer science as well as computing. The list in its raw form had 150 entries which were expanded to 252 words as there were many entries in the list with non-atomic values. For instance, the first entry with 'access' and 'memory access' was treated as two words viz. 'access' and 'memory access'. It is notable that 'memory access' is a MWP and was treated as a unigram with the form 'memory-access'. The snapshot of the expanded technical word list is presented in Table II. After the subtraction of 252 words from 1791 tokens, the resultant list had 1758 tokens. It is notable that only matching words were subtracted.

Finally, the POS-tagger [12] provided by the Stanford University was used for tagging the words in the paper titles. The resultant list had 2067 entries corresponding to 1758 tokens. The number of entries in the resultant list is more as the same token could be tagged with multiple parts of speech. As we were interested in the exhaustive coverage of the vocabulary, we considered all possible POS tags for a token.

Having created a POS-tagged token list of the database domain, the first part of the aim of creating an exhaustive vocabulary of the database domain, was achieved. For the second and last part, we targeted the creation of a POS-tagged lemmata list. This was achieved by lemmatizing the tokens in the list having 1758 tokens. It is noteworthy that we did not use the stemmer and used the Lemmatizer directly to obtain the lemma for each token rather than the non-lemma root for each token. The Lemmatizer [13] provided by the University of Copenhagen was used for this. The resultant list had 1530 lemmata.

The list with 1530 lemmata was further subjected to POS-tagging. The complete process was achieved through the use of multiple POS-taggers. Also, the different POS-taggers provided us with different sub-forms of POS tags but for simplicity the results were captured under the more common supersets. For instance, all entries from set {JJ (Adjective), JJR (Adjective, comparative), JJS (Adjective, superlative)} were considered to be just 'adjective' while entries from sets { NN (Noun, singular or mass), NNS (Noun, plural), NNP (Proper noun, singular), NNPS (Proper noun, plural)} and { VB (Verb, base form), VBD (Verb, past tense), VBG (Verb, gerund or present participle), VBN (Verb, past participle), VBP (Verb, non-3rd person singular present), VBZ (Verb, 3rd person singular present)} were considered to be 'noun' and 'verb' respectively.

Firstly, the Stanford University's POS-tagger [12] was used which resulted in 1504 entries corresponding to 1078 lemmata. The remaining unprocessed 452 lemmata were attempted to be POS-tagged using the Princeton University's POS-tagger [14]. This still resulted in 383 entries corresponding to only 261 additional lemmata. The remaining 191 remnant lemmata were manually POS-tagged. This resulted in 205 entries corresponding to 191 lemmata. Hence, the total number of entries corresponding to 1530 lemmata was 2092 in total. The summary of this data is presented in Table III.

TABLE I. A SNAPSHOT OF THE RAINBOW STOP WORD LIST [38]

Sr. No.	Stop Word
1	a
2	able
3	about
…	…
526	zero

TABLE II. A SNAPSHOT OF THE EXPANDED TECHNICAL WORD LIST FOR COMPUTER SCIENCE

Sr. No.	Technical Word
1	Access
2	access-time
3	Accumulator
…	…
252	Window

TABLE III. STATISTICS ON PROCESSING OF LEMMATA FOR POS USING DIFFERENT POS-TAGGERS

Sr. No.	POS-tagger	Lemma Count	POS Count
1	Stanford University POS-tagger [12]	1078	1504
2	Princeton University POS-tagger [14]	261	383
3	Manual POS-tagging (for remnant lemmata)	191	205
Total	3	1530	2092

Similar to the token POS-tagging case, for lemmata too there were multiple occurrences of a lemma having more than one POS-tag. Like the token POS-tagging case, in order to have an exhaustive coverage of the vocabulary of the technical domain of databases, we considered all possible POS tags for each lemma. The technical vocabulary of the database domain called DBTechVoc, which is a coined term, is formed by the POS-tagged token list and POS-tagged lemmata list.

IV. RESULTS AND DISCUSSION

The present research work was initiated with the motive of generating the technical vocabulary for the database domain. The titles of high-quality research papers in the field of database were believed to be the best source for populating the corpus. In order to make sure that no bias creeps in and also to assure that the basic terminology from the early days of development of databases as well as the latest terminology of the field of databases is covered, the duration of 45 years was considered for the present research work. Notably, this time period is not just long enough but also coinciding with the time period of evolution as well as proliferation of the field of databases. Also, it is both significant as well as relevant to consider the titles of research papers as something new in the field is first promulgated through a research paper and authors always include the important terms in the title of the research paper. It is with passage of time that those terms then become the part of the technical conversation of the field and thereby generating the technical field specific vocabulary.

TABLE IV. COMMON WORDS OF COMPUTER SCIENCE DOMAIN AND DATABASE DOMAIN

Sr. No.	Common Word	Sr. No.	Common Word
1	access	18	interface
2	allocation	19	interoperability
3	architecture	20	interpreter
4	backup	21	overhead
5	block	22	partition
6	buffer	23	pointer
7	cache	24	processor
8	capacity	25	protocol
9	disc	26	resolution
10	disk	27	retrieval
11	document	28	simulation
12	editor	29	software
13	error	30	statement
14	execution	31	storage
15	fragmentation	32	utility
16	hardware	33	window
17	instruction		

An important finding was obtained during the text processing with removal of stop words. It was observed that only 5.74% (or approx. 6%) of the tokens constituted the stop words. The same has been calculated using the formula: { [n(Tokens_with_SW) – n(Tokens_without_SW)] / n(Tokens_with_SW) } x 100 = Percentage of SW in Text; i.e. { [1900 – 1791] / 1900 } x 100 = 5.74 %. Here SW stands for Stop Words and n(entity) indicates the count of specified entity. This finding is in line with our assumption that the technical vocabulary of the database domain could be created from the titles of research papers as they contain more of important words rather than irrelevant words (like stop words). This holds true from multiple viewpoints of research, linguistics as well as statistics. Similarly, it was expected that a large number of words will be removed from the token list when computer science and computing domain technical word list will be considered. Actually, this step resulted in removal of just (1791-1758=) 33 common words. This means that there is only a { (33 / 1791) x 100 = 1.84% }, i.e. nearly 2% overlap of the technical lists of the domains of computer science and databases. The list of these removed common technical words is presented in Table IV. This finding is also very important and in line with our assumption that the technical word list of computer science domain will not be the same as the technical word list for the specific domain of databases.

Stemming was not used and directly lemmatization was used for the present research work to obtain the lemma for each token. Notably, this stage resulted in reduction of 13% entries from 1758 tokens to 1530 lemmata. This is important for the current context as it indicates the highly inflectional use of a few tokens by the researchers in the database domain. This is also important as it yielded a more refined vocabulary of the domain and hence let us meet the research objective.

TABLE V. DBTECHVOC (PART A): LIST OF TOKENS[A] AND CORRESPONDING POS

Sr. No.	Token	POS
1	abstract	noun
2	abstraction	noun
3	abstractions	noun
4	abstractions	verb
5	accelerating	noun
6	accelerating	verb
7	acceleration	noun
8	accesses	noun
9	accessibility	noun
10	account	noun
11	accuracy	noun
12	accurate	adjective
13	accurate	noun
14	achieving	noun
…	…	…
2065	xsketch	noun
2066	xsq	noun
2067	years	noun

[A]:Total unique tokens: 1758

Total unique POS: 5

TABLE VI. ANALYSIS OF FREQUENCIES OF POS TYPES OF TOKENS

Sr. No.	Token POS Type	Frequency	Share of POS Type (in %)
1	Noun	1341	64.88
2	Adjective	364	17.61
3	Verb	332	16.06
4	Adverb	25	1.21
5	Foreign Word (FW)	5	0.24
Total	5	2067	100

After following the various stages of methodology mentioned in section III, a final refined list of tokens was generated which was further POS-tagged. A snapshot of this list is presented in Table V. This table presents the glimpse of first 14 POS-tagged tokens and last 3 POS-tagged tokens from a total of 2067 POS-tagged tokens corresponding to 1758 unique tokens fortified with 5 unique POS. A summary on frequency of these unique 5 POS tags for this list is presented in Table VI. It can be observed from Table VI that nouns followed by adjectives constitute more than 82% of the total POS types.

Similar to the POS-tagged token list, another list for POS-tagged lemmata was also generated. A snapshot of this list is presented in Table VII. This table presents the glimpse of first 11 and last 8 POS-tagged lemmas out of a total of 1859 such POS-tagged lemmas corresponding to 1530 unique lemmas. A summary on frequency of the 5 unique POS tags found for this list (already presented in Table VII) is presented in Table VIII. It can be observed from Table VIII that the nouns and adjectives together constitute more than 80% of all the POS tags.

TABLE VII. DBTECHVOC (PART B): LIST OF LEMMATAA AND CORRESPONDING POS

Sr. No.	Lemma	POS
1	abstract	noun
2	abstraction	noun
3	acceleration	noun
4	access	noun
5	access	verb
6	accessibility	noun
7	account	noun
8	accuracy	noun
9	accurate	adjective
10	accurate	noun
11	achieve	verb
…	…	…
1852	xml	noun
1853	xpath	noun
1854	xqbe	noun
1855	xquery	noun
1856	xsd	noun
1857	xsketch	noun
1858	xsq	noun
1859	year	Noun

A:Total unique lemma: 1530

Total unique POS: 5

To summarize, Table V and Table VII present the POS-tagged token list and lemmata list respectively. The frequency break-up of the unique POS tags for Table V and Table VII is presented respectively in Table VI and Table VIII. The two lists viz. POS-tagged token list and POS-tagged lemmata list, together constitute the technical vocabulary of the database domain and have been addressed with a coined term DBTechVoc. Table V and Table VII represent the two parts, viz. A and B for DBTechVoc. The lists are presented in ascending order of the tokens and lemmata respectively. Similarly, the data in Table VI and Table VIII is sorted on the frequency of the POS-tag. Notably, both the tables ended up with same order of the POS-tags though their frequencies were different for the lists corresponding to tokens and lemmata. Fig. 2 presents the share (in units of percentage of the total count) of POS type for tokens and lemmata. It can be observed that there is no much difference between the breakup of POS types for tokens and lemmata. Notably, the number of adverbs, verbs and adjectives are more in case of lemmata list compared to those in the list of tokens.

TABLE VIII. ANALYSIS OF FREQUENCIES OF POS TYPES OF LEMMATA

Sr. No.	Lemmata POS Type	Frequency	Share of POS Type (in %)
1	Noun	1129	60.73
2	Adjective	365	19.63
3	Verb	329	17.70
4	Adverb	31	1.67
5	Foreign Word (FW)	5	0.27
Total	5	1859	100

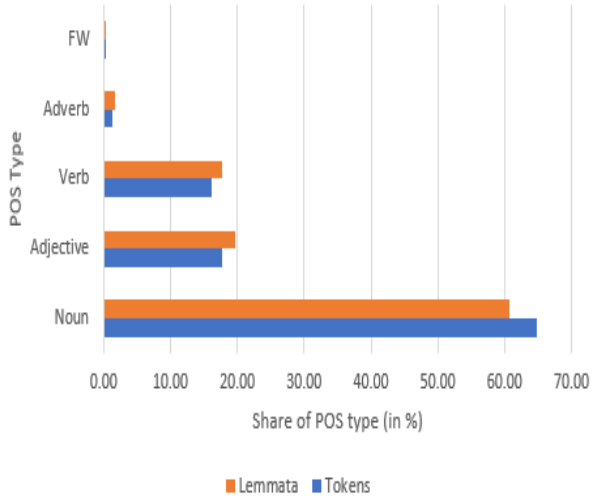


Fig. 2. Representation of Share (in %) of POS Types for Tokens and Lemmata.

In order to complement the vocabulary analysis of the database technical domain and to visualize the results of graphically, the word cloud, presented in Fig. 3, was generated. The word cloud was generated without stemming and lemmatization of the words though stop words were excluded from the list. Internal stop-word list was used during the execution of the code. The case of words was not considered and only unique words were considered for rendering through the cloud. In order to maintain the sanctity of data, the Multi-word Phrases (MWP) or the word formations with multiple words joined together with a hyphen like 'entity-relational', 'multi-valued' and 'grammar-based', just to name a few, were considered as-it-is without ignoring the hyphen. The number of words satisfying all these criteria was 1900 and out of these the cloud could accommodate the top 654 words. The frequency break-up of the remaining 1246 words which were not drawn through the cloud is given in Table IX.

As is clear from Fig. 3, the top most frequent word was 'database'. In order to further refine our analysis, the top 20 words were subjected to lemmatization. This resulted in the reduction of count and leading to 16 unique lemmata. These lemmata along with their corresponding Part-Of-Speech (POS) are shown in Table X. Notably, other than one verb and two adjectives, all other lemmata are nouns. This leads to an important inference that the authors tend to use more of nouns in the paper titles, at least for the research domain of databases.

TABLE IX. FREQUENCY BREAK-UP OF WORDS NOT DRAWN THROUGH THE CLOUD

Frequency of words	Number of words
1	940
2	175
3	111
4	7
5	10
6	3
Total	1246

Fig. 3. Word Cloud of the Most Frequent Terms in the Database Vocabulary Corpus Created for 45 Years.

TABLE X. Top 16 Lemmata and Corresponding POS

Sr. No.	Lemmata	POS
1	algorithm	Noun
2	analysis	Noun
3	approach	Noun
4	data	Noun
5	database	Noun
6	design	Noun
7	distribute	Verb
8	efficient	Adjective
9	information	Noun
10	language	Noun
11	model	Noun
12	processing	Noun
13	query	Noun
14	relational	Adjective
15	system	Noun
16	xml	Noun

As the proposed work is unique and first of its kind, its comparison as well as performance evaluation with respect to existing works is not feasible. However, the proposed work is better than the existing ones in terms of presenting a more specific vocabulary as well as better annotated vocabulary of the technical words of the database sub-domain of the computer science domain.

V. Conclusion, Limitations and Future Work

The present research work is the first formal attempt to create a technical vocabulary for the domain of databases. This vocabulary called DBTechVoc consists of a POS-tagged token list having 1758 multi-word phrase unigrams and a POS-tagged lemmata list having 1530 multi-word phrase unigrams. It is noteworthy that most of the Natural Language Processing (NLP) applications for generation of various word lists generally do not consider the multi-word phrases owing to the ease of processing that way. It is remarkable that as the present research work intended to create a technical vocabulary without the loss of semantic information of the technical phrases, the multi-word phrases have been well considered.

The various results and findings of the present research work are bound to have a good ripple effect for the researchers working in the same and similar fields. From the processing of more than 1000 research papers of last 45 years, it is concluded that the authors use 6% stop words in the titles of the research papers. Also, 13% of the words used for the research papers titles are inflectional forms of lemmata from a set consisting of tokens from the technical domain. There is a negligible overlap between the technical word lists for computer science domain and database domain. Also, based on perhaps first of its kind comparison between the frequency break-up of POS categories for tokens and frequency break-up of POS-categories of the corresponding lemmata of the tokens, it is concluded that the lemmatization results in increase in the number of adverbs, verbs and adjectives while reducing the number of nouns. Though the results reported here are for the technical domain of databases, they could be applied to other technical domains also as the period of 45 years and more than 1000 research papers is believed to be enough to normalize the values. All these results could be applied for analysis of and investigation on various works including the usage of these results as an aid to solve cases dealing with plagiarism as well as author attribution. This application may include research papers as well as other literary works, including touching on the areas of violation of copyrights and other intellectual property rights.

DBTechVoc itself could be used for various downstream tasks dealing with NLP of technical domains. DBTechVoc lists or the derived ones could also be used as a source of stop-words for some advanced applications dealing with the processing of this technical domain specific textual data, for instance, processing of social media reviews or opinions or comments on a particular topic dealing with the field of databases. The presented lists could also be used for generation of artificial language specific to the database domain. The lists could also be used for the readability analysis of the technical domains, particularly the databases. The proposed lists could also be used alongside the technical word list of the field of computer science in general as it happens to be the parent field of the domain of databases. Additionally, the lists could also be used for word-embeddings, Machine Translation Systems (MTS) and generation of a domain-specific technical WordNet.

One of the limitations of the present research work is that it presents the technical vocabulary of only the database domain. Also, though standard stop word list, technical word list, Lemmatizers and POS-taggers have been used, the results may differ if a different combination of these items is used. The findings, results and technical vocabulary presented here are all best reported as per the context and scope of the present research work. Though we believe that the proposed list tends to be exhaustive as on moment, it is notable that the field of database, like any other technical field, keeps on evolving and with passage of time, new words could be added to the domain. As future work, in addition to keeping the lists updated with appropriate versioning, we plan to consider the other parts of the published research papers like abstract, keywords, manuscript body, etc. for further fortifying the research methodology. Also, in addition to just the unigrams, bigrams, trigrams, etc. could also be considered for the vocabulary creation. Most importantly, with the measurement of semantic similarity between the tokens, we are working to generate a technical wordnet specifically for the database domain.

REFERENCES

[1] Liu N., Nation I.S.P. (1985), "Factors affecting guessing vocabulary in context", RELC Journal, 16(1):33–42.

[2] Dewan P., Gupta P. (2016), "Writing the Title, Abstract and Introduction: Looks Matter!", Indian Pediatrics, 53:235-241. Online: https://www.indianpediatrics.net/mar2016/mar-235-241.htm.

[3] Tullu M.S. (2019), "Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key", Saudi Journal of Anaesthesia, 13(1):S12-S17. doi: 10.4103/sja.SJA_685_18.

[4] Mack C. (2012), "How to write a good scientific paper:title, abstract, and keywords", Journal of Micro/Nanolithography, MEMS and MOEMS, 11(2):020101-1--020101-4.

[5] Karagel H., Karagel D.U. (2014), "Identification and importance of headings and key words in research in the framework of geography methodology", Procedia - Social and Behavioral Science, 120:356-364. doi: 10.1016/j.sbspro.2014.02.113.

[6] Soler V. (2007), "Writing titles in science: An exploratory study", English for Specific Purposes, 26:90–102. doi:10.1016/j.esp.2006.08.001.

[7] Hengl T., Gould M. (2002), "Rules of thumb for writing research articles", Enschede, pp.1-9. Online: https://webapps.itc.utwente.nl/librarywww/papers/hengl_rules.pdf.

[8] Webster J.J., Kit C. (1992), "Tokenization as the Initial Phase in NLP", in proceedings of COLING-92, pp. 1106-1110. Online: https://aclanthology.org/C92-4173.pdf.

[9] Kyle C. (1989), "Double, Triple, and Quadruple Bigrams," Word Ways, 22(3), art. 8. Online: https://digitalcommons.butler.edu/wordways/vol22/iss3/8.

[10] Akhmetov I., Pak A. Ualiyeva I., Gelbukh A. (2020), "Highly Language-Independent Word Lemmatization Using a Machine-Learning Classifier", Computación y Sistemas, 24(3):1353–1364. doi: 10.13053/CyS-24-3-3775.

[11] Maggini M. (n.d.), "Natural Language Processing Part 2: Part of Speech Tagging ", Teaching Slides, Department of Information Engineering and Mathematical Sciences, University of Siena, Italy. Online: https://www3.diism.unisi.it/~maggini/Teaching/TEL/slides%20EN/06%20-%20NLP%20-%20PoS%20Tagging.pdf.

[12] Stanford University (n.d.), "Part-Of-Speech (POS) Tagger", The Stanford Natural Language Processing Group. Online: http://nlp.stanford.edu:8080/parser/index.jsp.

[13] University of Copenhagen (n.d.), "CST's Part-Of-Speech tagger", Center for Language Technology. Online: https://cst.dk/online/pos_tagger/uk/.

[14] Princeton University (n.d.), "WordNet Search-3.1", WordNet: A Lexical Database for English. Online: http://wordnetweb.princeton.edu/perl/webwn.

[15] Smith S. (2019). (n.d.). Online: https://www.eapfoundation.com/vocab/other/lists/#thetable.

[16] West M. (1953), "A General Service List of English Words", London: Longman, Green and Co.

[17] Brezina V., Gablasova D. (2015), "Is There a Core General Vocabulary? Introducing the New General Service List", Applied Linguistics, 36(1):1–22. doi: 10.1093/applin/amt018.

[18] Browne C. (2013), "The New General Service List: Celebrating 60 years of Vocabulary Learning", The Language Teacher, 4(37):13–16.

[19] Gilner L., Morales F. (2008), "Elicitation and application of a phonetic description of the General Service List", System, 36(4):517–533.

[20] Gilner L. (2011), "A primer on the General Service List", Reading in a Foreign Language, 23(1):65–83. Online: https://files.eric.ed.gov/fulltext/EJ926367.pdf.

[21] Nation P., Waring R. (2004), "Vocabulary size, text coverage and word lists". Online: https://web.archive.org/web/20080111133710/http://www.wordhacker.com/.

[22] Kaur J., Saini J.R. (2015), "POS Word Class based Categorization of Gurmukhi Language Stemmed Stop Words", in proceedings of Springer International Conference on ICT for Intelligent Systems (ICTIS-2015), Ahmedabad, India, 51(2):3-10. doi: 10.1007/978-3-319-30927-9_1.

[23] Kaur J., Saini J.R. (2016), "Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle ", in proceedings of Symposium on ACM Women in Research (ACM-WIR-2016), Indore, India, 01188:32-37. doi: 10.1145/2909067.2909073.

[24] Rakholia R.M., Saini J.R. (2017), "A Rule-based Approach to Identify Stop Words for Gujarati Language", in proceedings of The 5th Springer International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2016), Bhubaneshwar, India, 515:797-806. doi: 10.1007/978-981-10-3153-3_79.

[25] Rakholia R.M., Saini J.R. (2016), "Lexical Classes Based Stop Words Categorization for Gujarati Language", in proceedings of 2nd IEEE International Conference on Advances in Computing, Communication & Automation (ICACCA-2016), Bareilly, India, pp. 1-5. doi: 10.1109/ICACCAF.2016.7749005.

[26] Hancioglu N., Neufeld S., Eldridge J. (2008), "Through the looking glass and into the land of lexico-grammar", English for Specific Purposes, 27(4):459-479. doi: 10.1016/j.esp.2008.08.001.

[27] Coxhead A. (2000), "A new Academic Word List", TESOL Quarterly, 34(2):213–238.

[28] Billurog˘lu A., Neufeld S. (2005), "The Bare Necessities in Lexis: A new perspective on vocabulary profiling". Online: http://lextutor.ca/vp/BNL_Rationale.doc.

[29] Bakaric M.B., Babic N., Matetic M. (2021), "Application-based Evaluation of Automatic Terminology Extraction", International Journal of Advanced Computer Science and Applications, 12(1):18-27. doi: 10.14569/IJACSA.2021.0120103.

[30] Choy M. (2012), "Effective Listings of Function Stop words for Twitter", International Journal of Advanced Computer Science and Applications, 3(6):8-11. doi: 10.14569/IJACSA.2012.030602.

[31] Raulji J.K., Saini J.R. (2017), "Generating Stopword List for Sanskrit Language", in proceedings of 7th IEEE International Advance Computing Conference (IACC-2017), Hyderabad, India, pp. 799-802. doi: 10.1109/IACC.2017.0164.

[32] Raulji J.K., Saini J.R. (2020), "Sanskrit Stopword Analysis through Morphological Analyzer and its Gujarati Equivalent for MT System", in proceedings of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, India, 93:427-433. doi: 10.1007/978-981-15-0630-7_42.

[33] Venugopal G., Saini J.R., Dhanya P. (2020), "Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List", International Journal of Advanced Computer Science and Applications, 11(1):233-239. doi: 10.14569/IJACSA.2020.0110130.

[34] Saini J.R., Rakholia R.M. (2016), "On Continent and Script-wise Divisions-based Statistical Measures for Stop-words Lists of International Languages", Procedia Computer Science, 89:313-319. doi: 10.1016/j.procs.2016.06.076.

[35] (n.d.) "Database Terminology – A Dictionary of the Top Database Terms". Online: https://raima.com/database-terminology/.

[36] Fayaza M.S.F., Farhath F.F. (2021), "Towards Stopwords Identification in Tamil Text Clustering", International Journal of Advanced Computer Science and Applications, 12(12):524-529. doi: 10.14569/IJACSA.2021.0121267.

[37] The ACM Digital Library (2022), "ACM Transactions on Database Systems", The Association for Computing Machinery. Online: https://dl.acm.org/journal/tods.

[38] Shuson N. (2022), "StopWords list based on Rainbow statistical text". Online: https://gist.github.com/shuson/b3051fae05b312360a18.

[39] Saed H., Hussein R.F., Haider A.S., Al-Salman S., Odeh I.M. (2022), "Establishing a COVID-19 lemmatized word list for journalists and ESP learners", Indonesian Journal of Applied Linguistics, 11(3): 577-588. doi: 10.17509/ijal.v11i3.37103.

[40] Das S.R., Donini M., Zafar M.B., He J., Kenthapadi K. (2022), "FinLex: An effective use of word embeddings for financial lexicon generation", The Journal of Finance and Data Science, 8:1-11. doi: 10.1016/j.jfds.2021.10.001.

[41] Ahsanuddin M., Hanafi Y., Basthomi Y., Taufiqurrahman F., Bukhori H.A., Samodra J., Widiati U., Wijayati P.H. (2022), "Building a corpus-based academic vocabulary list of four languages", Pegem Journal of Education and Instruction, 12(1):159–167. doi: 10.47750/pegegog.12.01.15.

[42] Joensuu J. (2022), "Culinary List Form in the Experimental Poetry of 1960s Finland: Literary Menus and Recipes", in: Barton R.A., Böckling J., Link S., Rüggemeier A. (eds) Forms of List-Making: Epistemic, Literary, and Visual Enumeration, Palgrave Macmillan, Cham. doi: 10.1007/978-3-030-76970-3_9.

[43] Wingrove P. (2022), "Academic lexical coverage in TED talks and academic lectures", English for Specific Purposes, 65:79-94. doi: 10.1016/j.esp.2021.09.004.

[44] Alasmary A. (2022), "Academic lexical bundles in graduate-level math texts: A corpus-based expert-approved list", Language Teaching Research, 26(1):99-123. doi: 10.1177/1362168819877306.

Smart Monitoring System for Chronic Kidney Disease Patients based on Fuzzy Logic and IoT

Govind Maniam[1], Jahariah Sampe[2]*
Institute of Microengineering and Nanoelectronics
Universiti Kebangsaan Malaysia (UKM)
43600 Bangi, Selangor
Malaysia

Rosmina Jaafar[3], Mohd Faisal Ibrahim[4]
Department of Electrical, Electronics and Systems
Engineering, Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia (UKM)
43600 Bangi, Selangor, Malaysia

***Abstract*—A Chronic Kidney Disease (CKD) monitoring system was proposed for early detection of cardiovascular disease (CVD) and anemia using Fuzzy Logic. To determine the heart rate and blood oxygen saturation, the proposed model was simulated using MATLAB and Simulink to handle ECG and PPG inputs. The Pan-Tompkins method was used to determine the heart rate, while the Takuo Aoyagi algorithm was used to assess blood oxygen saturation levels. The findings show that the ECG recorded using the CKD model has all of the characteristics of a typical ECG wave cycle, but with reduced signal degradation in the 0.8–1.3mV region. The heart rate signal processing yielded findings between 78 and 83 beats per minute is within the range of the supplied heart rate. Takuo Aoyagi's pulse oximeter simulation generated the same findings. For real-time verification, the proposed model was implemented in hardware using ESP8266 32-bit microcontroller with IoT integration via Wireless Fidelity for data storage and monitoring. In comparison with the Fuzzy Logic simulation done on MATLAB and Simulink, the CKD monitoring device has 100% accuracy in patient status detection. The CKD monitoring system has an overall accuracy of 99% in comparison with a commercial fingertip pulse oximeter.**

***Keywords*—Anemia; cardiovascular disease (CVD); fuzzy logic; healthcare; internet of things**

I. Introduction

Healthcare monitoring systems or e-Health systems are devices that use a wireless sensor network (WSN) to observe severe or chronic diseases in humans [1] In this era, there are many smart watches out in the market that claim to track the condition of the body accurately. However, these smartwatches could not be used to diagnose a medical condition. It can only give alerts on an abnormal vital sign [2]. In this case, we have to use a proper medical device to monitor the vital signs. Some examples of medical devices are heart rate monitors (HRM), pulse oximeters, electrocardiogram (ECG), blood pressure monitors, thermometer, etc. Monitoring vital signs play an important role in healthcare monitoring systems. The vital signs of patients in intensive care unit (ICU) are also observed using healthcare monitoring systems [3]. Chronic disease patients require constant monitoring. In 2005, chronic disease fatality rates increased, with a total death count of more than 58 million people worldwide [4]. According to a recent study from Malaysia's National Renal Registry, for ten year's there has been an increase in new dialysis patients from 4,606 to 8,431 from 2008 to 2018 [5].

As a result, healthcare monitoring systems serve an important role in monitoring patients' vital signs and early detection of prevailing diseases. However, the high cost of equipment has been one of the drawbacks of a healthcare monitoring system. Treatment and monitoring chronic diseases cost a lot of money in both low-income and high-income nations urging the need for low-cost healthcare solutions [6]. Diseases that are caused by chronic diseases may be avoided if patients' vitals are monitored. Chronic Kidney Disease (CKD) is reported to outnumber other chronic illnesses including cardiovascular disease (CVD) and anemia in this scenario [7]. Due to insufficient erythropoietin hormones in CKD patients, previous studies reveal that the prevalence of CVD [8] and anemia [9] are very significant. Monitoring the electrocardiogram (ECG), heart rate, and blood oxygen saturation level (SpO_2) can help in early prevention. Contaminants in ECG signals are commonly divided into the following groups. Power line interference, electrode pop or contact noise, patient–electrode motion artefacts, electromyographic (EMG) noise, and baseline wandering are all examples of these problems [10]. These pollutants cause ECG readings to be inaccurate, making it harder to diagnose the heart's activity.

The integration of an Artificial Intelligence in a healthcare monitoring device will improve the decision-making process. Fuzzy Logic is known to be a form of artificial intelligence where the approach to computing is based on "degrees of truth" rather than the usual "true or false". This allows human-like reasoning to take place to identify pathologies in a person.

The design of a sensor interface controller for early detection of anemia and CVD in CKD patients with the aid of artificial intelligence is the focus of this article. The implementation of Internet of Things (IoT) will further enhance the device with transmitting and storing data via Cloud Computing. ECG signal generation, heart rate detection, SpO_2 and patient condition are all done with MATLAB and Simulink. For early diagnosis of CVD and anemias, a Fuzzy Logic Interface (FIS) is implemented. The Fuzzy Logic Toolbox graphical user interface (GUI) from MATLAB is used to simulate the FIS. The suggested method is implemented in hardware using ESP8266 microcontroller for real-time verification.

*Corresponding Author.

II. PPG Signal and ECG Signal for SpO_2 and Heart Rate Detection

Pulse oximeter sensors generate photoplethysmogram (PPG) signals, which have both an AC and DC component, as seen in Fig. 1. The AC component refers to the Pulsatile Arterial Blood Absorption, which describes how light is absorbed by blood circulating through arteries [11].

The SpO_2 is determined using an equation pioneered in the 1970s by Takuo Aoyagi [12]. Aoyagi tweaked Wood's plot of red and infrared light haemoglobin density to design a SpO_2 sensor that can detect the necessity for artificial ventilation [13]. The SpO_2 is determined by sensing light attenuation through a haemoglobin absorptive medium. The Beer-Law, Lambert's which connects the concentration of a solute in a solvent to the absorption of light passing through the solution [14], is used to do this. Equation (1) represents the relationship between the concentration of the solute and the absorption of light.

$$I_{\lambda out} = I_{\lambda in} 10^{-\varepsilon \lambda c l} \quad (1)$$

Where $I_{\lambda out}$ is the intensity of transmitted light, $I_{\lambda in}$ is the intensity of incident light, λ is the wavelength of light, ε is the extinction coefficient of solute, c is the concentration of solute and finally l is the length of path that the incident light travels through.

The SpO_2 is estimated using red (visible) light and the infrared light (IR) where both lights contain AC and DC component using equations (2) and (3).

$$R = \frac{AC_{Red}/DC_{Red}}{AC_{IR}/DC_{IR}} \quad (2)$$

$$SpO_2 = 110 - 25R \quad (3)$$

Where, R is the AC to DC ratio of the red light divided by the AC to DC ratio of the infrared (IR). AC_{Red} is the pulsating AC component of the red light. AC_{IR} is the pulsating AC component of the infrared light. DC_{Red} is the DC component of the red light and DC_{IR} is the DC component of the infrared light. The normal range of SpO_2 levels is 96% – 100%.

Fig. 2 shows the relationship between the amplitude of red light (R) and infrared light (IR) to the SpO_2 and red light to infrared light ratio (R/IR) [15]. When the amplitude of both the red and infrared light is the same it gives an R/IR ratio of 1.0 which produces an SpO_2 of 85%. Where else, -3.4 R/IR ratio produces an SpO_2 of 0% and 0.43 R/IR produces an SpO_2 of 100%. These amplitude ratio of red light to infrared light are used to model the simulation blocks.

The purpose of an ECG machine is to collect electrical signals from the heart's activity. The information about the heart's activity will be presented in a waveform pattern by the ECG machine. Cardiologists use 12 lead ECG as the gold standard instrument to monitor the heart functions to detect the changes from normal heart rhythm. Findings such as abnormalities from visual inspections of the ECG waveforms will be the basis for necessary further heart examinations such as angiograms. The regular cycle of the ECG waveform representing the heart's activity is shown in Fig. 3. The P wave indicates atrial contractions to transfer blood into the ventricle, the QRS complex indicates ventricular contraction, and the T wave represents ventricular repolarization [16].

Instead of utilizing the PPG data from the pulse oximeter sensor, the ECG signals from the ECG sensor were used to determine the heart rate. The R peak is the point on the ECG signal with the highest amplitude that could be clearly identified. The QRS complex has been filtered off in earlier research to make it easier to detect R peaks [17]. Because the R-R interval is clearly distinguishable in the ECG signal, the R peaks may be utilized to window it. The R-R interval, as illustrated in Fig. 3, can be used to determine a person's heart rate. The heart rate (HR) in beats per minute (BPM) is calculated using the number of R peaks recorded in one minute [18]. Equation (4) can be used to calculate heart rate.

$$HR = \frac{60{,}000ms/min}{R\text{–}R\ interval(ms)} \quad (4)$$

The R-R peaks is measured using millisecond, hence millisecond is applied. As a result, one minute will be divided between R-R peaks. If the R-R intervals are 800ms, for example, (60,000ms/min)(800)ms = 75 BPM.

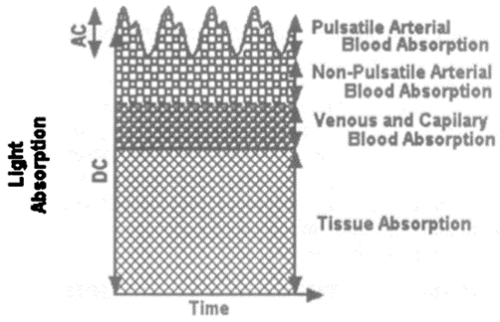


Fig. 1. Signal Components of PPG Signal [11].

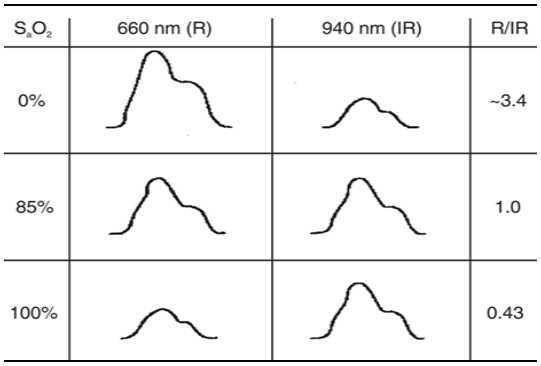


Fig. 2. Relationship between the Amplitude of R and IR to the SpO_2 and R to IR Ratio [15].

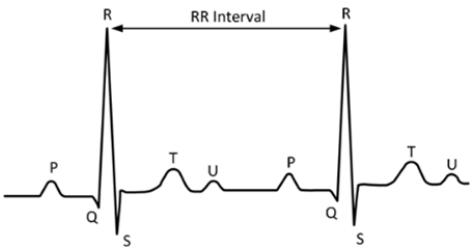


Fig. 3. Regular ECG Wave Cycle and R-R Peaks [10].

III. Design and Development

In order to measure and monitor the vital signs of CKD patients, the design, development, and simulation of the CKD monitoring system are done using MATLAB and Simulink. Fig. 4 shows the proposed CKD monitoring system integrated with IoT. The system consists of seven components which are patient, smart sensors, device, connectivity, cloud platform, application, and user. The first component is the patient who will be monitored. Smart sensors consist of a pulse oximeter sensor and an ECG sensor. The third component is device also known as 'thing'. The brain to the system is the microcontroller which will process signals from the smart sensors. The outputs are sent to the fuzzy logic interface for decision-making on the patient's conditions. The results will then be displayed locally on an OLED screen. The fourth component is connectivity via Wireless Fidelity (Wi-Fi). The Wi-Fi will be used to link the device to the fifth component of the system which is the cloud platform. The cloud platform consists of the cloud computing and cloud storage. Data from the device will be stored in the cloud as well as sent to the sixth component, application. The application consists of smartphones and the web. This system will allow users to view the results using smartphones and web browsing. The last component of the framework is the user. The user consists of caretakers as well as healthcare professionals. This system allows healthcare professionals to personally monitor the patient all the time. The patient's data will be stored on the cloud for easier access to the patient's history.

Simulations have been done on MATLAB and Simulink for signal processing of PPG signals and ECG signals. The PPG signals were processed to obtain the blood oxygen level readings while the ECG signals were processed to obtain the filtered ECG and heart rate reading.

The monitoring system consists of algorithms and techniques to determine the heart rate and SpO_2 levels. Fig. 5 shows the proposed CKD monitoring system to monitor the condition of CKD patients. The model consists of signals generated as inputs imitating the pulse oximeter sensor and the ECG sensor. The model also has a subsystem block as a microprocessor that contains the algorithms for measuring heart rate and SpO_2. Finally, scopes and displays are set as outputs to display the results.

The ECG signal block contains 5 presets of ECG signals which generate different heart rates when selected. The presets are named very low, low, normal, high, and very high that generate 45 bpm, 60bpm, 78 - 83bpm, 160bpm, and 220 bpm respectively. Before being shown on the ECG signal scope, the produced signals will go via a sample rate converter to match the output sample rate [19]. The sample rate converter has a tolerance of 0.01 and a sample output rate of 200Hz. The Pan-Tompkins method, which includes a Band pass Filter, Differentiator Filter, Moving Average Window, and QRS Peak Detection [20], will be used to estimate heart rate from ECG data. The sample rate converter block is used to match the source sampling rate to the output sampling rate [21]–[23]. A 198Hz two-sided bandwidth of interest was used to transform the sample rate. A Band pass Filter, Differentiator Filter, Moving Average Window, QRS Peak Detection, and Unbuffer make up the ECG signal processing [24]. The ECG signal processor's role is to filter the ECG signal so that the patient's heart rate may be determined. The band pass filter [25]–[28] is a mixture of a high-pass and a low-pass filter. The band pass filter eliminates noise from muscle movements, breathing fluctuation, and baseline wander. The band pass filter in this model is set to correct any attenuation of the QRS complex and eliminate artifacts from the heart's motion. Equation (5) shows the transfer function of a second-order low-pass filter with a high cutoff frequency of around 11Hz.

$$H(z) = \frac{(1-z^{-6})^2}{(1-z^{-1})^2} \quad (5)$$

The transfer function of a high-pass filter with a low cutoff frequency of nearly 5Hz is shown in Equation (6).

$$H(z) = \frac{-\frac{1}{32}+z^{-16}+z^{-17}+\frac{1}{32}z^{-32}}{1-z^{-1}} \quad (6)$$

A full-band differentiator filter is used by the differentiator filter to isolate all of the frequency components in the input signal [24]–[26]. An equiripple Finite Impulse Response (FIR) filter architecture is used to construct this block. This block's filter order is set to 51, and the maximal passband ripple is left at preset. The Pan-Tompkins algorithm's differentiator filter is based on equation (7), with a 2 sample output signal delay.

$$H(z) = \frac{1}{10}(-2^{-2} - z^{-1} + z + 2z^2) \quad (7)$$

A 'Discrete FIR Filter' is used to construct the moving average window block. This block has been used to transfer the data that has been obtained one by one. This block primarily determines the window size. Equation (8) represents the result of the Pan-Tompkins algorithm, where N is the window width, which varies depending on the size of samples.

$$y(nT) = \frac{1}{N}[x(nT-(N-1)T) + x(nT-(N-2)T + x(xnT) \quad (8)$$

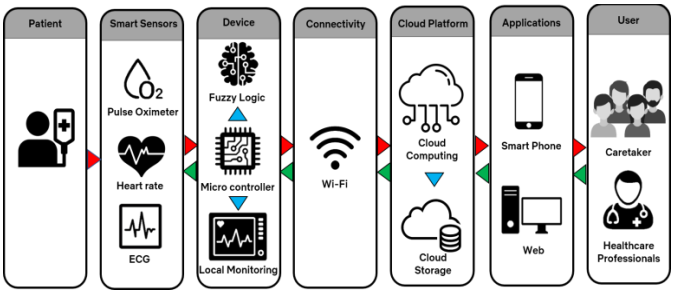


Fig. 4. Proposed CKD Monitoring System Integrated with IoT.

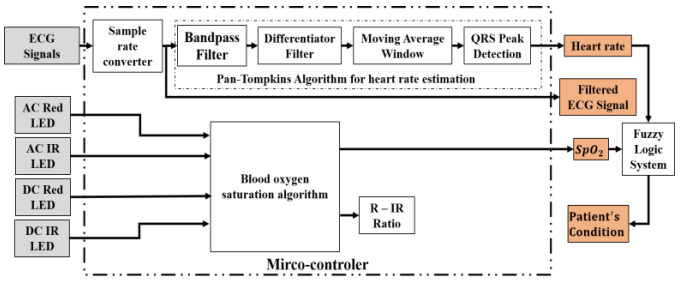


Fig. 5. Block Diagram of Proposed CKD Monitoring System.

To identify R peaks and estimate the patient's heart rate, the QRS peak detection block is used. The R peaks and the ECG signal threshold can be used to estimate the patient's heart rate. The threshold and R peaks' parameters threshold are first declared. The R peak amplitudes are programmed to detect a range of 0.055–0.075mV with a width of more than 10.01ms. Any signal that does not fall within the specified range will be ignored and deemed noise. The threshold is calculated using the average noise peak and the mean estimations of average R peaks with 8 samples. The R-R peaks will be analyzed to measure the heart rate using the identified R peaks. The following equation (9) and (10) was used in the Pan-Tompkins algorithm to detect R peaks.

$$Peak > ThrSig \rightarrow SigLev = \frac{1}{8}peak + \frac{7}{8}SigLev \quad (9)$$

$$ThrNoise = \frac{1}{2}ThrSig \quad (10)$$

Where Peak is the total peak. ThrSig is the signal peak threshold. SigLev is the signal peak running estimate. ThrNoise is the noise peak threshold.

The SpO_2 was measured using Equation (2). The PPG signals from a pulse oximeter consist of an AC and DC component from Infrared LED and a Red LED [29]. In total there are four sources of signals produced. The generated signals are represented by the source generator AC RLED, AC IRLED, DC IRLED, and DC RLED. The AC signals are built using a Repeating Signal Generator. A repeating series of integers provided in a table of time-value pairs is produced as the output. Time values should be growing consistently. The amplitudes of the AC IRLED and AC RLED were set at different points to result in a difference in the R value generated [30]. The R value determines the estimation of the SpO_2 levels of the patient. The DC IRLED and DC RLED were generated using a pulse generator which will generate DC signals at different amplitudes. The inputs were connected to a MATLAB function block which is coded using Equation (1) and Equation (2) [31], [32]. The results of the R value and SpO_2 on a display block. The simulation was done with different inputs applied to the source generator. The outputs were then recorded to verify the function of the simulation model.

The Fuzzy Logic Interface System (FIS) is used to predict the deterioration of vital signs for early detection of abnormalities in the patient's body by using heart rate and SpO_2 as the parameters. The Fuzzy Logic Toolbox™ graphical user interface (GUI) from MATLAB is used to run the simulation. The toolbox includes the Fuzzy Logic Designer, Membership Function Editor, Fuzzy Rules, Rule Viewer and the Surface Viewer.

IV. Simulation Results

The Simulation provides results of the ECG, heart rate, and the SpO_2. In comparison with a typical ECG theoretical cycle wave, the ECG generated demonstrates the patient's normal sinus with a resting heart rate range of 78 and 83 BPM. The heart rate at the input was set between 78 and 83 beats per minute, and as shown in Fig. 6, the heart rate is 82 beats per minute. Confirming that the Pan-Tompkins algorithm used to estimate the HR in this simulation is acceptable.

Fig. 7 displays the pre-recorded ECG signals that were used to synthesize the ECG signals. In comparison with a regular ECG theoretical cycle wave, the ECG signal clearly demonstrates the patient's normal sinus with a resting heart rate of 78–83 BPM. The amplitude of the signals is indicated on the y-axis in mV, and the time is indicated on the y-axis in seconds. The P, R, and T waves have peak amplitudes of 0.79 – 0.88mV, 1.01 – 1.28mV, and 0.95 – 1.05mV, sequentially. The T wave has a larger peak amplitude than the P wave, as predicted.

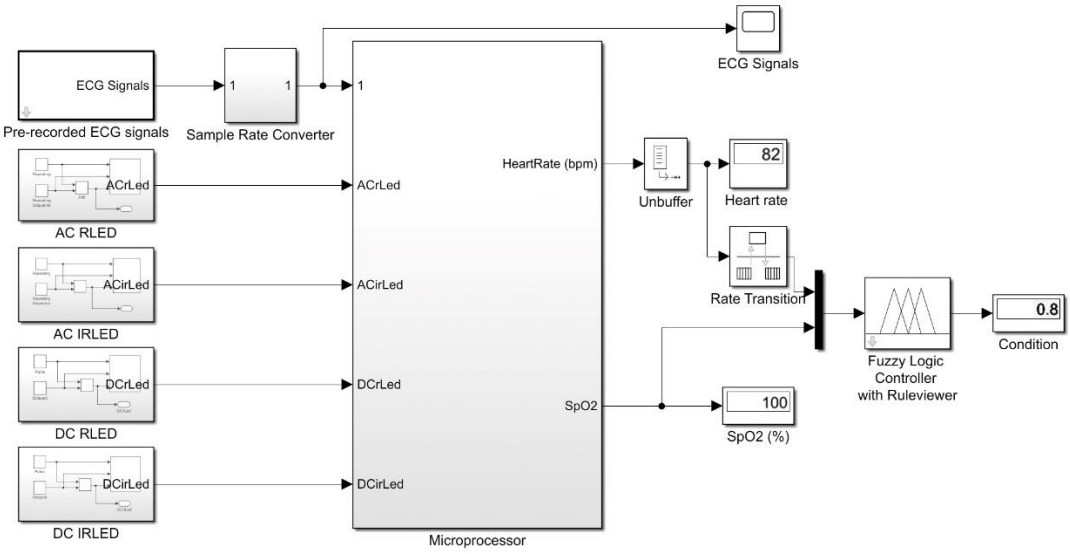


Fig. 6. Simulink Model of CKD Monitoring System.

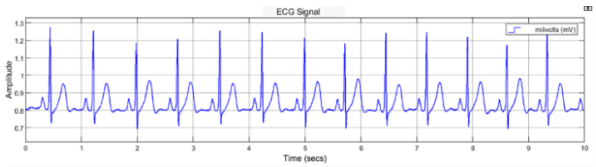


Fig. 7. Pre-recorded ECG Signal Generation.

The ECG waveforms are shown in Fig. 8 at different phases of digital signal processing for R peak identification. In comparison to the Pan-Tompkins algorithm theoretical findings, the results are similar at every stage. The amplitude in mV is indicated on the y-axis, while the time in seconds is indicated on the x-axis. The output of the ECG signals after being filtered using a band pass filter is shown in Fig. 8(a). Fig. 7 shows that the peak amplitudes of the P, T, and U waves are relatively low than those of the signals in Fig. 8. (a). Using a low-pass and high-pass filter method, the band pass filter has filtered out the high and low signals, leading the P, T, and U waves to have a reduced amplitude. Fig. 8(b) displays the result of the next phase, which involves filtering ECG signals using a differentiator filter. The QRS complex has a larger amplitude than the P, T, and U waves, which have a smaller amplitude. The amplitude of the R peak has declined from 0.3mV to 0.12mV. On the other hand, the amplitudes of the P, T, and U peaks are in the range of 0.01–0.04mV. In Fig. 8(a), it's clearly seen that this filter has also eliminated the negative numbers. The output of the moving average filter can be seen in Fig. 8(c). The moving average window generates a signal that contains information about the QRS complex's slope and breadth. The last stage in signal processing for R peak detection is shown in Fig. 8(d). After applying the adaptive thresholds, the processed data display a stream of pulses indicating the positions of the QRS complexes. At the same time, the P, T, and U are totally filtered out by the moving average window. The amplitude of these pulses is between 0.05 and 0.07 millivolts.

Fig. 6 illustrates the Simulink model in action when the ECG source is set to 78–83 BPM. As a result, the heart rate shown on the "Display" Simulink block is 82 BPM, demonstrating the validity of the Pan-Tompkins algorithm used in this simulation. Fig. 8(d) shows the R peaks that were used to compute the heart rate. A second is equal to 60Hz, hence if more than one R peak is observed in a second, the signals have a frequency greater than 60Hz. Fig. 8(d) shows that within a minute more than one R peak is recorded. This indicates that the heart rate of this ECG signal is more than 60 beats per minute. As a result, the Simulink block appears to be capable of processing the ECG data in order to retrieve the R peaks and heart rate. Or else, the Simulink model is obliged to have an inaccuracy if a heart rate range of 78 – 83 BPM was not presented.

The inputs of the AC RLED and AC IRLED are manipulated to obtain different SpO_2 readings. Table I shows the relationship between the inputs, R value, and the SpO_2 reading. The relationship between the R value and the SpO_2. The R to IR ratio was manipulated by changing the values of the AC RLED and the AC IRLED inputs. When the AC RLED has a higher value than the AC IRLED the R to IR ratio is high causing the SpO_2 value to be low, vice versa when the AC IRLED has a higher value than the AC RLED. These results are similar to the pulse oximeter design study by Jubran (1996) and prove the validity of the simulation model in different SpO_2 levels in the patient's body.

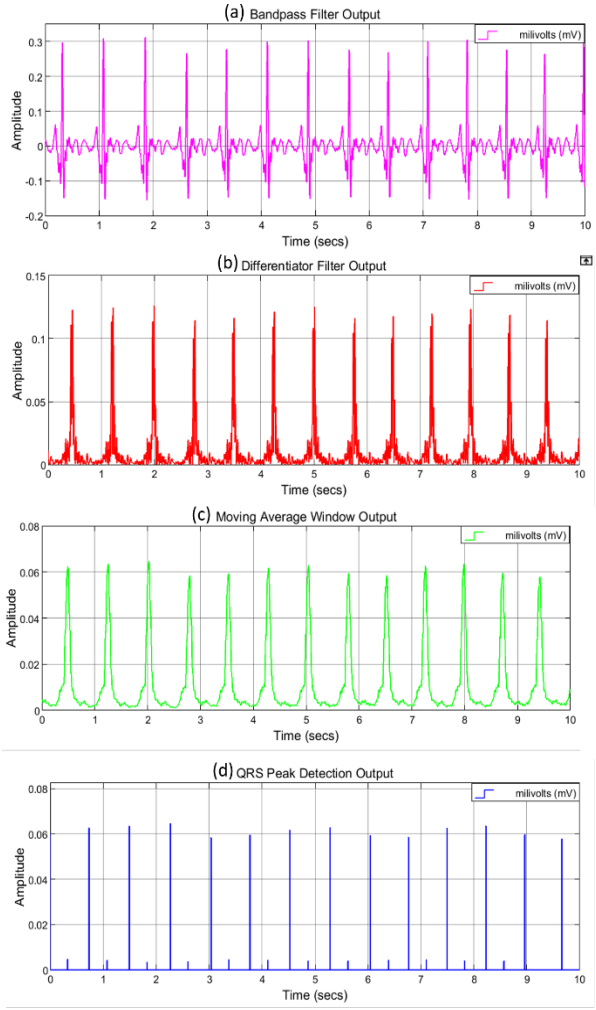


Fig. 8. Phases in Digital Signal Processing and QRS Peak Detection a. Bandpass Filter Output. b. Differentiator Filter Output. c. Moving Average Window Output. d. QRS Peak Detection Output.

TABLE I. R TO IR RATIO AND ITS RELATIONSHIP TO SpO_2 ESTIMATION

AC RLED	AC IRLED	R to R Ratio	SpO_2 Estimation
2.0	6.0	0.33	101.67
2.0	5.0	0.40	100.00
2.0	4.0	0.50	97.50
2.0	3.0	0.67	93.33
2.0	2.0	1.00	85.00
3.0	2.0	1.50	72.50
4.0	2.0	2.00	60.00
5.0	2.0	2.50	47.50
6.0	2.0	3.00	35.00
7.0	2.0	3.50	22.50
8.0	2.0	4.00	10.00
9.0	2.0	4.50	-2.50

V. Abnormalities Detection using Fuzzy Logic

A Mamdani fuzzy logic-based detection system is designed as shown in Fig. 9. The system receives two inputs that are heart rate and SpO_2 readings and provides one output which is the patient's condition. MATLAB software is used to build the fuzzy logic system. Next, the Membership Function Editor is used to determine the shape for each membership function that is associated with the declared variables. The "trapmf" Membership Function that provides the trapezoidal-shaped relationship between a crisp variable and its corresponding fuzzy values was used. Fig. 10 and 11 show the declared membership function for the inputs. The heart rate has 5 membership functions while the SpO_2 has 3 membership functions. Fig. 12 shows the patient condition membership functions which determine the condition of the patients to be normal, abnormal, or critical. The centroid defuzzification method is done by using a closed-form of membership functions. This method returns the crisp value that corresponds to the fuzzy set's center of area. Following the declaration of membership functions for the inputs and output, the fuzzy rule-base is used to establish a specific set of output functions depending on certain specified inputs, as determined by medical specialists using ground truth base in Table II [33]–[35]. Fig. 13 shows the rules set up in the Rule Editor. In total, 15 rules are resulting in 3 possible outcomes depending on the parameter of inputs. The outcome of the patient's condition is normal when the vital signs namely heart rate and SpO_2 are in the range of 70 – 100 BPM and 96 – 100%, respectively.

TABLE II. Ground Truth Parameters used for Fuzzy RuleBase

Parameters	Range	Interpretation
Heart rate	0 – 50 50 – 70 70 – 100 100 – 160 160 – 230	Critically Low Bradycardia Normal Tachycardia Critically High
SpO_2	30 – 70 70 – 95 96 – 100	Critically Low Low Normal

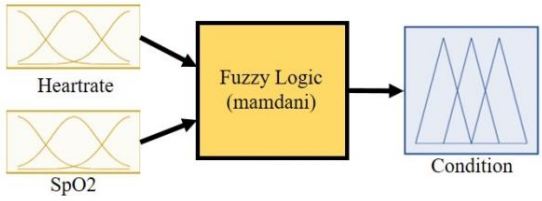


Fig. 9. The Designed Fuzzy Logic.

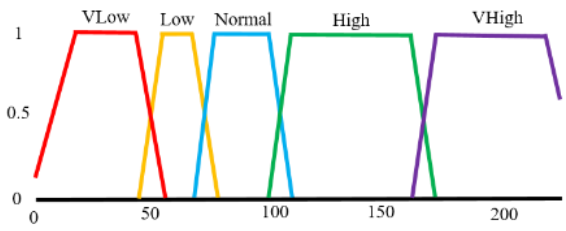


Fig. 10. Heart Rate Membership Function.

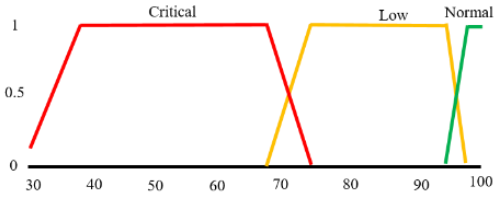


Fig. 11. SpO_2 Membership Function.

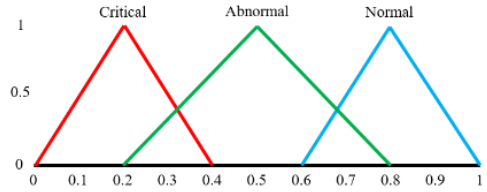


Fig. 12. Patient Condition.

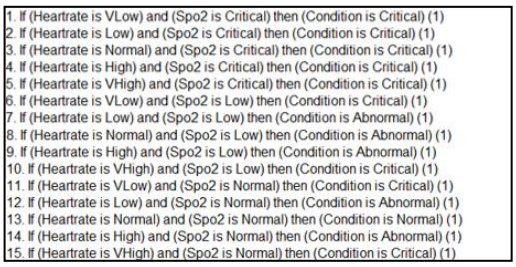


Fig. 13. Fuzzy Logic IF-THEN Rule base Membership Function.

The result of the fuzzy logic system can be simulated in MATLAB using the Rule Viewer as shown in Fig. 14. The Rule Viewer can be used to identify whether the expected parameters are obtained for the given inputs. The stability of the system and the accuracy can be estimated with the help of the diagram. For example in the simulation done in Fig. 6, the heart rate is set at 82 BPM and the SpO_2 is set at 100% and the output shows the patient is in normal condition. The fuzzy logic system was tested in all possible outcomes to verify the stability and accuracy of the rules. The Surface Viewer is utilized to see how one of the output is affected by one or more inputs. It constructs and plots a system output surface map, as illustrated in Fig. 15. With a successful simulation, the Fuzzy Logic variables, parameters, membership functions, and rules were coded into ESP8266 32-bit microcontroller to make a smart CKD monitoring system.

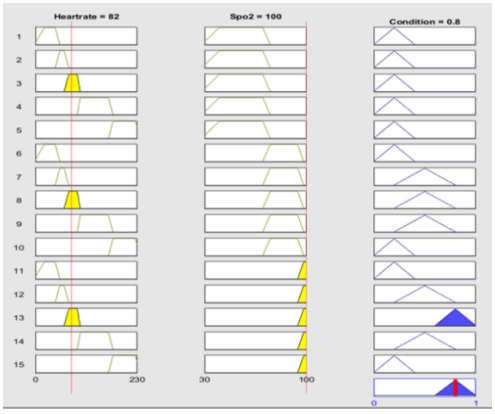


Fig. 14. Rule Viewer.

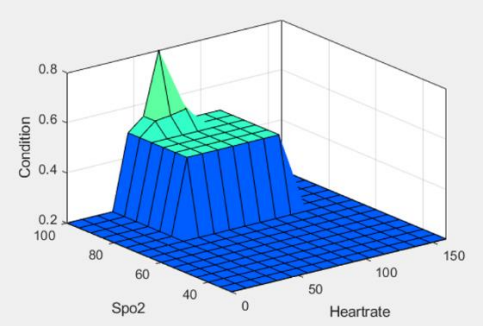


Fig. 15. Surface Viewer.

The Fuzzy Logic Controller was interfaced in the Simulink model to simulate the function. Table III shows the results from the tested parameters. The condition of the patient is displayed as "Normal" when the heart rate and SpO_2 are only in the normal range.

Table IV shows the previous studies conducted by other researchers to monitor vital parameters in the past few years. The comparisons were focused on the types of microcontrollers used, the presence of simulation, the parameters that were measured, the inclusion of an abnormalities detection method, and the accuracy of the device. Many of the studies utilizing the Arduino microcontroller had high accuracy. In this study, Fuzzy Logic will be used for the early detection of anemia and CVD in CKD patients. The results of the simulation, hardware and a typical medical device were compared to prove the accuracy of the simulation and hardware developed.

TABLE III. RESULTS FROM FUZZY LOGIC CONTROLLER SIMULATION

Heart rate (BPM)	**SpO_2 (%)**	**Condition**
60 – 85	96 – 100	Normal
86 – 130	96 – 100	Abnormal
131 – 200	96 – 100	Critical
30 – 60	96 – 100	Abnormal
60 – 85	70 – 95	Abnormal
60 – 85	30 – 69	Critical

TABLE IV. RELATED RESEARCHES ON HEALTHCARE MONITORING SYSTEMS

Ref	**Microcontroller**	**Simulation**	**Parameters**	**Abnormalities Detection Method**	**IoT Implementations**	**Accuracy**
[36]	Arduino UNO	None	• Heart rate • SpO_2 • Temperature	None	Cloud and Andriod App	98%
[37]	ATMega328p	Fluke ProSim 8 Vital Sign Simulator	• ECG • Heart rate • SpO_2 • Temperature	Coding on Arduino	Cloud	99%
[38]	Arduino Nano	None	• ECG	Cloud	Andriod App	-
[39]	8-bit Atmel Microcontroller	None	• Heart rate • SpO	None	Bluetooth to PC	-
Authors	ESP8266 32-bit	MATLAB and Simulink	• ECG • Heart rate • SpO_2	Fuzzy Logic	Cloud	99%

VI. HARDWARE IMPLEMENTATIONS

The successful simulation of the CKD monitoring system will be implemented into hardware for real-time verification. The hardware consists of a MAX30105 Pulse Oximeter, AD8232 ECG Sensor, and AD8232 ECG electrodes as the inputs. The Arduino based ESP8266 32-bit was used as the microcomputer and the OLED screen was used as the output to display the results. The Arduino based ESP8266 board has a built-in Wi-Fi feature that does not need an external Wi-Fi module making this system to be compact. The built-in Wi-Fi module has an IEEE 802.11 b/g/n that uses various frequencies including, but not limited to, 2.4 GHz, 5 GHz, 6 GHz, and 60 GHz frequency bands. In the case of noisy settings, the AD8232 has a signal conditioning block that can retrieve, enhance, and filter weak bio-potential signals. This implies that signal contamination from motion artifacts or remote electrode placement can be minimized. The AD8232 comes with 3-lead ECG electrodes that plug into a 3mm audio jack. The MAX30105 is an integrated particle-sensing module that can be used to produce PPG signals from the arterial pulse. The MAX30105 communicates through a standard I^2C compatible interface. This makes it easier for the microcontroller to process information with a simple circuit. The module uses a red light with a wavelength of 680nm and infrared light with a wavelength of 880nm. It also comes with a built-in digital filter and an analog to digital (ADC) signal converter.

Fig. 16 shows the wiring diagram of the CKD monitoring system. The MAX30105 Pulse Oximeter and the OLED Screen are interfaced via the I^2C module at the A4 and A5 pins. The AD8232 ECG monitor is interfaced to the analog pin A0 of the ESP8266 while the Lo+ and Lo- of the ECG module are interfaced to digital pins D5 and D6 respectively. The AD8232 ECG leads are connected to the 3mm audio jack of the AD8232 ECG module. The ESP8266 is powered using a 9V battery. The ESP8266 microcontroller is programmed using the Arduino's Integrated Development Environment (IDE) via C programming language. The libraries of the sensors, displays, and fuzzy logic systems were installed before programming.

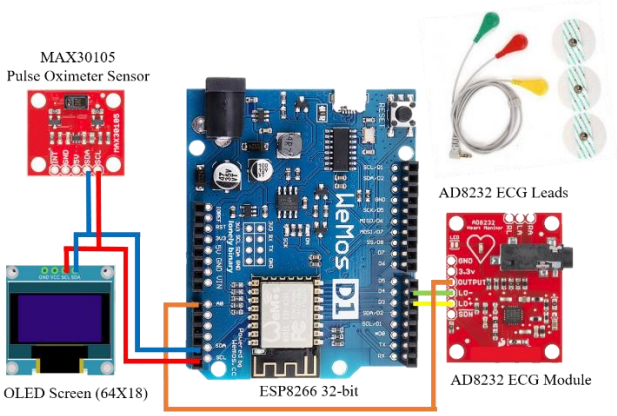


Fig. 16. Wiring Diagram of the Implemented Hardware using ESP8266.

The Arduino based ESP8266 microcontroller was programmed using the Arduino IDE. The program was written using C++ language. Fig. 17 shows the flowchart of the program that was coded into the ESP8266 microcontroller. The microcontroller initializes the libraries that will be used to run the program. The initialization includes the AD8232, MAX30105, SDD106 OLED, Fuzzy Logic, and Wi-Fi. Then, the connection to the internet will be attempted. With a successful internet connection, the Cloud Platform server will be done. Once connected the smart sensors will get the data from the patient. The ECG, heart rate, and SpO_2 will be obtained by processing the ECG signals and PPG signals. The obtained data will be sent simultaneously to the Fuzzy Logic interface and displayed on the OLED screen. The results from the Fuzzy Logic will be displayed on the same OLED as well. Finally, all the obtained data will be sent to the cloud platform via a Wi-Fi connection. If the device is not turn off, it will repeat the loop by obtaining data from the smart sensors.

The prototype of the CKD monitoring system hardware is shown in Fig. 18. The result of SpO_2 estimation was compared to the value obtained by a commercial fingertip pulse oximeter as shown in Fig. 19 to validate the measurement. The developed CKD monitoring system was tested on a normal person. The ECG leads were placed on the chest during a supine position [40] according to Einthoven's triangle for ECG lead placement as shown in Fig. 20.

The ECG leads are color coded where the green (left leg = LL) serves as the reference electrode, red, is for the left arm (LA), and yellow is for the right arm (RA). The RA and LA leads are placed below the right and left clavicle respectively. The LL lead is placed below the left rib bones. The ECG leads must not be placed on the bones to avoid obstructing the signals and it can cause noise. The left index finger is placed on the MAX 30105 pulse oximeter while the right index finger is put inside the commercial pulse oximeter. After a few seconds, the results were displayed. On the OLED panel, the ECG waveforms obtained by the AD8232 ECG sensor were vividly presented. In comparison to a typical ECG theoretical cycle wave, the ECG includes all of the elements of a regular sinus. Along with the output from the fuzzy logic controller, the heart rate and SpO_2 were clearly shown. The condition of the patient is displayed as "Nor" as the heart rate and SpO_2 were in the normal range of 70 BPM and 98%, respectively.

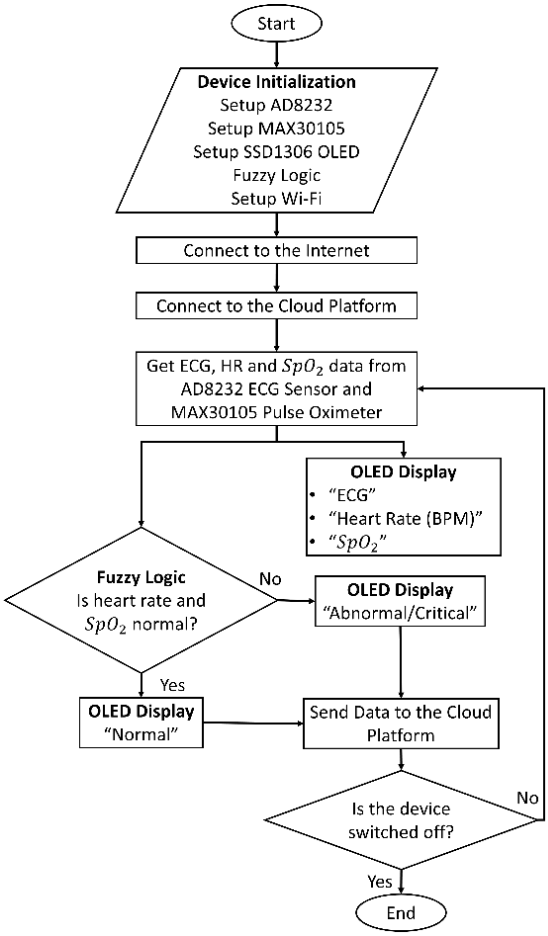


Fig. 17. Flowchart of the Microcontroller Program.

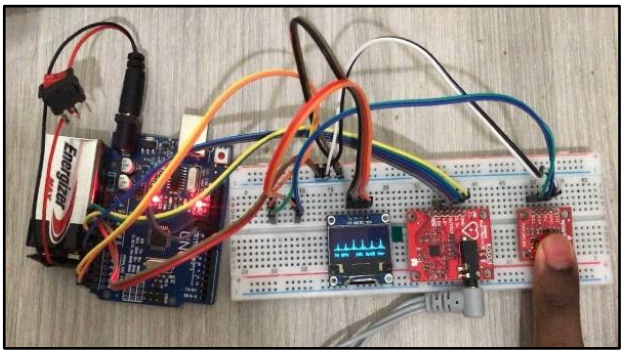


Fig. 18. CKD Monitoring System Hardware.

Fig. 19. Fingertip Pulse Oximeter.

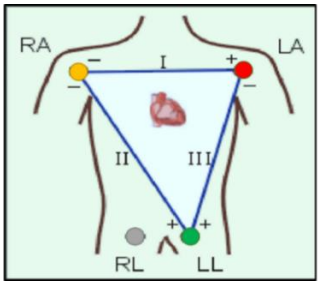


Fig. 20. Einthoven's Triangle for ECG Lead Placement [40].

The cloud platform was set up using the Arduino Cloud IoT. The Arduino Cloud IoT is an application programming interface (API) that provides services for developers to configure, program, and connect to other devices. In this project, the ESP8266 microcontroller was connected to the internet via Wi-Fi to send the information to the server containing dashboards and storage. Fig. 21 shows the interface of the smartphone application. The first page of the application displays dashboard consisting of the patient's ID. When the patient ID is selected in the dashboard the second page opens. The second page displays the SpO_2, heart rate, and the patient's condition. The SpO_2 and heart rate of the patient displays 97% and 69BPM respectively simultaneous to the results displayed in the local OLED screen. The same results can also be seen on the website and logged in to the cloud storage.

Table V shows the results from comparing the developed CKD monitoring system to the commercial fingertip pulse oximeter. The heart rate has a deviation of 1 – 2 BPM while the SpO_2 has a deviation of 1%. The fuzzy logic system programmed into the ESP8266 was compared to the results simulated in MATLAB and Simulink. The cloud application represents the readings obtained from the smartphone. Results show that there is no deviation in the outcome of the patient's condition, thus proving 100% accuracy of the CKD monitoring system IoT-based. The prototype CKD monitoring system has an overall accuracy of 99%.

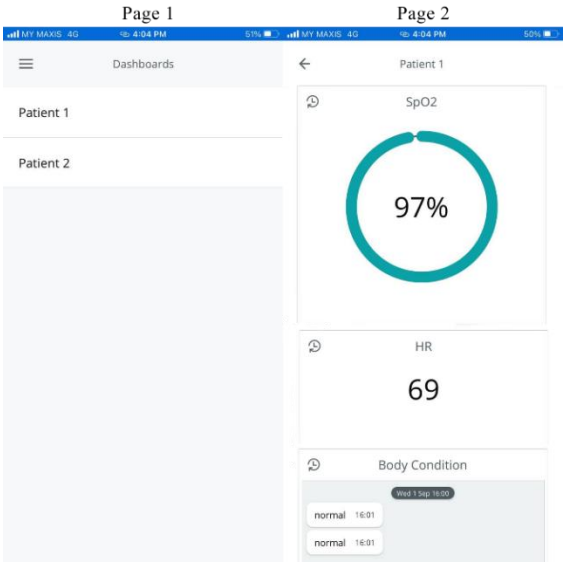


Fig. 21. IoT Smartphone Application Interface.

TABLE V. ACCURACY OF THE DEVELOPED IOT BASED CKD MONITORING SYSTEM

Parameters	Accuracy	Deviation
Heart rate	99.2%	1 – 2 BPM
SpO_2	99.4%	1 %
Fuzzy Logic	100%	None
Cloud Platform	100%	None

The proposed device has minor limitations that can be improved in future works. The biosensors, OLED display, and microcontroller are powered by a 9V battery that consumes a lot of power and prevents the device from being used for an extended period. Low powered biosensors and displays can be added or developed as an upgrade.

VII. CONCLUSION

A model of a simulation was designed to imitate the functionality of a low-cost monitoring system for early detection of CVD and anemia for CKD patients. The ECG signals produced in the simulation were similar to the theoretical cycle in the range of 0.8 – 1.3mV with different heart rate range settings. The simulation done on SpO_2 to yield similar results to the Takuo Aoyagi algorithm. The hardware implementation verified the prototype CKD monitoring system in real-time by displaying the ECG, heart rate, and SpO_2 along with the patient's condition with 99% accuracy when compared with a commercial fingertip pulse oximeter. The fuzzy logic coded into the ESP8266 32-bit microcontroller was accurate by 100% relative to the simulation done on MATLAB and Simulink. The suggested system contributes to the decrease of fatalities from chronic kidney disease patients by monitoring important parameters including ECG, HR, and SpO_2 with the inclusion of decision making using Fuzzy Logic.

ACKNOWLEDGMENT

This work is funded by Ministry of Education Malaysia under the grants (TRGS/1/2019/UKM/01/4/3) and (FRGS/1/2018/TK04/UKM/02/1).

REFERENCES

[1] N. H. Mohd Yunus, J. Sampe, J. Yunas, A. Pawi, and Z. A. Rhazali, "MEMS Based Antenna of Energy Harvester for Wireless Sensor Node," Microsyst. Technol., vol. 26, no. 9, pp. 2785–2792, 2020, doi: 10.1007/s00542-020-04842-5.

[2] H. N. Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, "A Review on Wearable Photoplethysmography Sensors and Their Potential Future Applications in Health Care," Int. J. Biosens. dna Bioelectron, vol. 176, no. 3, pp. 139–148, 2019, doi: 10.15406/ijbsbe.2018.04.00125.A.

[3] S. Birnbaum, "Pulse Oximetry Identifying its Applications, Coding, and Reimbursement," Chest, vol. 135, no. 3, pp. 838–841, 2009, doi: 10.1378/chest.07-3127.

[4] WHO, "Preventing Chronic Diseases :A Vital Investment," 2015.

[5] G. B. Leong and L. D. Guat, Dialysis in Malaysia. 2018.

[6] OECD/WHO, Health at a Glance: Asia/Pacific 2020, vol. 6011, no. 24312. 2020.

[7] M. E. Stauffer and T. Fan, "Prevalence Of Anemia in Chronic Kidney Disease in the United States," PLoS One, vol. 9, no. 1, pp. 2–5, 2014, doi: 10.1371/journal.pone.0084943.

[8] M. Volpe et al., "Blood Levels Of Erythropoietin in Congestive Heart Failure and Correlation with Clinical, Hemodynamic, and Hormonal

Profiles," Am. J. Cardiol., vol. 74, no. 5, pp. 468–473, 1994, doi: 10.1016/0002-9149(94)90905-9.

[9] G. Sunder-Plassmann and W. H. Hörl, "Effect of Erythropoietin on Cardiovascular Diseases," Am. J. Kidney Dis., vol. 38, no. 4 SUPPL. 1, pp. 20–25, 2001, doi: 10.1053/ajkd.2001.27391.

[10] N. Djermanova, M. Marinov, B. Ganev, S. Tabakov, and G. Nikolov, "LabVIEW Based ECG Signal Acquisition and Analysis," 2016 25th Int. Sci. Conf. Electron. 2016, 2016, doi: 10.1109/ET.2016.7753471.

[11] T. L. Rusch, R. Sankar, and J. E. Scharf, "Signal Processing Methods for Pulse Oximetry," Comput. Biol. Med., vol. 26, no. 2, pp. 143–159, 1996, doi: 10.1016/0010-4825(95)00049-6.

[12] J. W. Severinghaus, "Takuo Aoyagi: Discovery of Pulse Oximetry," Anesth. Analg., vol. 105, no. SUPPL. 6, p. S1, 2007, doi: 10.1213/01.ane.0000269514.31660.09.

[13] G. A. Millikan, "The Oximeter, an Instrument for Measuring Continuously the Oxygen Saturation of Arterial Blood in Man," Rev. Sci. Instrum., vol. 13, no. 10, pp. 434–444, 1942, doi: 10.1063/1.1769941.

[14] H. P. S. Saini, "Simulink Based Modelling of a Pulse Oximeter," California State University, Northridge Simulink, 2013.

[15] A. Jubran, "Pulse Oximetry," Crit Care, vol. 3, no. 2, pp. 605–608, 1999, doi: 10.1542/pir.2018-0123.

[16] C. Tso, G. M. Currie, D. Gilmore, and H. Kiat, "Electrocardiography: A Technologist's Guide to Interpretation," J. Nucl. Med. Technol., vol. 43, no. 4, pp. 247–252, 2015, doi: 10.2967/jnmt.115.163501.

[17] T. P. Utomo, N. Nuryani, and Darmanto, "QRS Peak Detection for Heart Rate Monitoring on Android Smartphone," J. Phys. Conf. Ser., vol. 909, no. 1, 2017, doi: 10.1088/1742-6596/909/1/012006.

[18] P. F. Shahina Begum, Mobyen Uddin Ahmed, Physiological Sensor Signals Analysis to Represent Cases in a Case-Based Diagnostic System. Springer, 2013.

[19] M. A. Albrni, M. Faseehuddin, J. Sampe, and S. H. M. Ali, "Novel VDBA Based Universal Filter Topologies with Minimum Passive Components," J. Eng. Res., vol. 9, no. 3, pp. 110–130, 2021, doi: 10.36909/jer.v9i3B.8781.

[20] N. V. T. Nguyen, L. D. Tran, and T. Van Huynh, "Detect QRS Complex in ECG," Proc. 2017 12th IEEE Conf. Ind. Electron. Appl. ICIEA 2017, vol. 2018-Febru, no. 3, pp. 2022–2027, 2018, doi: 10.1109/ICIEA.2017.8283170.

[21] N. A. Nayan, R. Jaafar, and N. S. Risman, "Development of Respiratory Rate Estimation Technique using Electrocardiogram and Photoplethysmogram for Continuous Health Monitoring," Bull. Electr. Eng. Informatics, vol. 7, no. 3, pp. 487–494, 2018, doi: 10.11591/eei.v7i3.1244.

[22] M. Z. Suboh, R. Jaafar, N. A. Nayan, and N. H. Harun, "ECG-based Detection and Prediction Models of Sudden Cardiac Death: Current Performances and New Perspectives on Signal Processing Techniques," Int. J. online Biomed. Eng., vol. 15, no. 15, pp. 110–126, 2019, doi: 10.3991/ijoe.v15i15.11688.

[23] M. Z. Suboh, R. Jaafar, N. A. Nayan, and N. H. Harun, "Shannon Energy Application for Detection of ECG R-peak using Bandpass Filter and Stockwell Transform Methods," Adv. Electr. Comput. Eng., vol. 20, no. 3, pp. 41–48, 2020, doi: 10.4316/AECE.2020.03005.

[24] G. Maniam, J. Sampe, A. A. Hamzah, M. Faseehuddin, and Noorhidayah, "Biosensor Interface Controller for Chronic Kidney Disease Monitoring Using Internet of Things (IoT)," J. Phys. Conf. Ser., vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012110.

[25] F. Mohammad, J. Sampe, S. Shireen, and S. Hamid Md Ali, "Minimum Passive Components Based Lossy and Lossless Inductor Simulators Employing A New Active Block," AEU - Int. J. Electron. Commun., vol. 82, pp. 226–240, 2017, doi: 10.1016/j.aeue.2017.08.046.

[26] M. Faseehuddin, N. Herencsar, M. A. Albrni, and J. Sampe, "Electronically Tunable Mixed-Mode Universal Filter Employing A Single Active Block and A Minimum Number of Passive Components," Appl. Sci., vol. 11, no. 1, pp. 1–26, 2021, doi: 10.3390/app11010055.

[27] M. Ibharim, A. Albrni, F. Mohammad, N. Herencsar, J. Sampe, and S. H. Ali, "Novel Electronically Tunable Biquadratic Mixed- Mode Universal Filter Capable of Operating in MISO and SIMO Configurations," J. Microelectron. Electron. Components Mater., vol. 50, no. 3, pp. 189–203, 2020.

[28] J. Sampe, M. Faseehuddin, and S. H. M. Ali, "Design of Ultra-Low Voltage CCII Utilizing Level Shifting Technique and A Dual Mode Multifunction Universal Filter as an Application," J. Eng. Res., vol. 6, no. 2, pp. 155–175, 2018.

[29] A. Mukherjea, P. Chaudhury, A. Karkun, S. Ghosh, and S. Bhowmick, "Synthesis of PPG Waveform Using PSPICE and Simulink Model," Proc. 3rd Int. Conf. 2019 Devices Integr. Circuit, DevIC 2019, pp. 428–432, 2019, doi: 10.1109/DEVIC.2019.8783684.

[30] K. Pilt, K. Meigas, J. Lass, and M. Rosmann, "Signal Processing Methods for PPG Module to Increase Signal Quality," IFMBE Proc., vol. 16, no. 1, pp. 434–437, 2007, doi: 10.1007/978-3-540-73044-6_111.

[31] M. Shokouhian, R. C. S. Morling, and I. Kale, "Simulink Based Behavioural Modelling of A Pulse Oximeter for Deployment in Rapid Development, Prototyping and Verification," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 10, no. 2, pp. 5566–5569, 2012, doi: 10.1109/EMBC.2012.6347255.

[32] M. Shokouhian, R. C. S. Morling, and I. Kale, "Low Cost MATLAB-Based Pulse Oximeter for Deployment in Research and Development Applications," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, pp. 1740–1743, 2013, doi: 10.1109/EMBC.2013.6609856.

[33] D. Shimbo, "Ambulatory Blood-Pressure Monitoring," no. June, 2014, doi: 10.1056/NEJMra060433.

[34] D. M. Nathan et al., "Medical Management of Hyperglycemia in Type 2 Diabetes: A Consensus Algorithm for the Initiation and Adjustment of Therapy," Diabetes Care, vol. 32, no. 1, pp. 193–203, 2009, doi: 10.2337/dc08-9025.

[35] L. Wilkins, ECG Interpretation Made Incredibly Easy!: Fifth edition. 2012.

[36] M. M. Ali, S. Haxha, M. M. Alam, C. Nwibor, and M. Sakel, "Design of Internet of Things (IoT) and Android Based Low Cost Health Monitoring Embedded System Wearable Sensor for Measuring SpO2, Heart Rate and Body Temperature Simultaneously," Wirel. Pers. Commun., vol. 111, no. 4, pp. 2449–2463, 2020, doi: 10.1007/s11277-019-06995-7.

[37] M. A. Yusof, S. X. Fung, W. L. Low, C. W. Lim, and Y. W. Hau, "Miniaturized And Portable Home-Based Vital Sign Monitor Design with Android Mobile Application," Int. J. Integr. Eng., vol. 11, no. 3, pp. 10–22, 2019, doi: 10.30880/ijie.2019.11.03.002.

[38] S. S. Sylvester, E. L. M. Su, C. F. Yeong, and F. K. Che Harun, "Miniaturized and Wearable Electrocardiogram (ECG) Device with Wireless Transmission," J. Telecommun. Electron. Comput. Eng., vol. 9, no. 3–9, pp. 15–19, 2017.

[39] S. P. Rekha Chandra R,. Safeer K. P, "Design and Development of Miniaturized Pulse Oximeter for Continuous Spo2 and HR Monitoring with Wireless Technology," Int. J. New Technol. Res., vol. 1, no. 1, p. 263706, 2015.

[40] R. Yadav, S. Vashisth, and A. K. Salhan, "Real Time Acquisition and Analysis of ECG signals using MATLAB," Int. J. Adv. Eng. Sci. Technol., vol. 2, no. August, pp. 190–195, 2016.

Trust Management in Industrial Internet of Things using a Trusted E-Lithe Protocol

Ahmed Motmi[1], Samah Alhazmi[2]*, Ahmed Abu-Khadrah[3], Mousa AL-Akhras[4], Fuad Alhosban[5]
Computer Science Department, College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia[1, 2, 3, 4]
Computer Information Systems Department, King Abdullah II School for Information Technology, The University of Jordan, Amman 11942, Jordan[4]
Computer Information Science Department, Faculty of Computer Information Systems, Higher Colleges of Technology, UAE[5]

***Abstract*—The IoT has gained significant recognition from research and industrial communities over the last decade. The concept of Industrial IoT (IIoT) has emerged to improve industrial processes and reduce downtime or breach in secure communication. If automated, industrial applications can make the implementation process more convenient, it also helps increase productivity, but an external attacker may cause distortion to the process, which could cause much damage. Thus, a trust management technique is proposed for securing IIoT. The transition of the Internet to IoT and for industrial applications to IIoT leads to numerous changes in the communication processes. This transition was initiated by wireless sensor networks that have unattended wireless topologies and were comprised due to the nature of their resource-constrained nodes. In order to protect the sensitivity of transmitted information, the security protocol uses the Datagram Transport Layer Security (DTLS) mandated by Secure Constrained Application Protocol (CoAP). However, DTLS was designed for powerful devices and needed strong support for industrial applications connected through high-bandwidth links. In the proposed trust management system, machine learning algorithms are used with an elastic slide window to handle bigger data and reduce the strain of massive communication. The proposed method detected on and off attacks on nodes, malicious nodes, healthy nodes, and broken nodes. This identification is necessary to check if a particular node could be trusted or not. The proposed technique successfully predicted 97% of nodes' behavior faster than other machine learning algorithms.**

***Keywords*—IoT; industrial; IIoT; trust management; E-lithe; secure communication; internet of things; CoAP; datagram transport layer security**

I. Introduction

Humans have been analyzing their surrounding physical environment for a thousand years, including identifying additional vital elements for maintaining the balance like measuring temperature, distance, and time. Initially, rudimentary methods were focused on using references like sun's position and body part sizes. With the technological advancement, the measurement units were standardized over time, the first-ever mechanical unit to offer exact physical measures appeared and were named sensors. With the electronic revolution in the silicon age, the method has been more precise for calculating and labeled as electronic sensors. United States Army introduced communication via sensors in the 1950s [1]. The research was initiated by Silverstein and was known as Sound Surveillance System. It was an intelligence-based project launched for detecting Soviet Submarines in the Pacific and Atlantic Oceans [2].

With time these sensors were improved for using trust-based communication through wireless network systems. The physical variables can now be modified through an artifact known as an actuator. The Wireless Sensor Networks (WSNs) are optimized thoroughly by incorporating sophisticated mechanisms and actuators, developing into Wireless Sensor and Actuator Network. Each node of the Wireless Sensor and actuator network can be turned into an Internet of Things (IoT) device using internet protocol [3].

IoT is conceptualized in this research as "An IoT device is an embedded system which is resource-constrained but has the capability of performing well-defined tasks like networking, signal processing, and sensing. It is powered by batteries and offers wireless communication capabilities" [2].

The concept of IoT has maximized the interoperability of devices. The connection and communication between devices have been facilitated but securing the connection is still questionable. The proposal for implementing IoT in computer connection and big data calculations is not new; however, the scope of the problem changes when implementing trust factors within the communication of Industrial IoT (IIoT) [4].

IoT is used for improvising domestic applications and has also been focused on innovating industrial applications. The research focuses on cyber security and trusted communication between industrial applications. Though IIoT offers quality domestic application uses, complex structures are needed to implement advanced communication techniques within industrial applications.

IIoT is carried out between hundreds of devices among hundreds or possibly thousands of devices connected to the same wireless network. It can create scalability issues, and an even larger amount of data transferred needs security and safe transmission without data theft and intrusion [5, 6]. Thus, the characteristic of interoperability needs to be controlled through improvising the security feature within IIoT. The efficiency of this technique may seem questionable considering implementing security features for a massive network that needs to be executed and maintained. Along with the

*Corresponding Author.

robustness and scalability requirements, the focus needs to be given to deploying fine-grained access control mechanisms [7].

According to Cisco and Gartner, currently, there are six billion users connected with the IoT devices; this number is increasing exponentially. It has been claimed by [7] that it is expected that IoT for industrial applications will help in improving security features and will offer great potential for research in this field in the coming years.

However, IoT has been associated with issues like resource-constraint devices with limited processing capabilities and their memory. The overhead is considered a technological barrier in this domain. The use of standard protocols only increased overhead delays and energy consumption [8]. The delays break communication connectivity, and overhead delays affect energy life. Both these factors are not acceptable for the efficient operation of any application. Thus, due to these issues, the questions stated below are developed to help direct research for bringing better prospects in deploying IoT in industrial applications and maintaining the trust factor [7].

1) Is the implementation of trusted communication for the Industrial Internet of Things feasible?

a) Are there any benefits of making IoT the baseline technology for IIoT Trust management?

b) If the performance impact is reduced, is it possible to increase interoperability?

c) Interoperability is the desired application of these services but increasing the possible number of inter-connections raises the chances of connecting many malicious users.

2) While maintaining performance, how can access exposed IoT nodes be controlled?

3) How can trust management be implemented while maintaining zero-configuration achieved for an IoT node?

To be able to answer these questions, a detailed analysis over energy, time consumption, and implementing trust factor in IIoT along with studying memory footprints and communication overheads, detailed analysis is conducted in coming sections of this research with the help of milestones depicted in Fig. 1 [7].

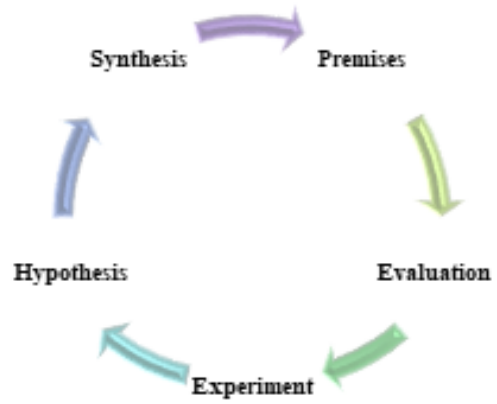


Fig. 1. Research Methodology [7].

The methodology described for this research is based on an iterative process depicted in Fig. 1. The iterative process needs a real-life problem. When the perimeters are targeted for the problems, it leads to the formulation of initial research questions. With the increasing knowledge, the depth of the problem may be well understood and lead to additional research questions. The iteration is expected to continue until the evaluation is successfully achieved [9].

The research will be divided into three stages. The first step is to examine the available research on network security, IoT technologies, Wireless Network Systems and authentication for IIoT, and the security protocols for lightweight key management through the preliminary study and literature survey. The second phase will be the theoretical design for resource-constrained sensor nodes in WSN and IoT using lightweight security solutions. The primary aspects of these proposed solutions were end-to-end (E2E) key management, device authentication, and communication in secure groups. Simulations, estimations, and real-time implementations in the third phase will evaluate the proposed solutions [10]. This research is aimed to evaluate through qualitative and quantitative research methods. Data will be collected from primary and secondary sources.

IoT requires collaboration from different research backgrounds and industries as a multi-disciplinary field. As industries are the major resource-generating entities, they need continuous improvement with evolving technology due to more human reliance and being operated manually; industries have been unable to operate optimally. This research proposes that IoT collaborate with industrial applications to produce an efficient IIoT model. All devices are expected to be automated and communicate through the Internet.

This research investigates, proposes, and analyzes efficient IoT technologies that will enable cutting-edge IoT networks to implement and update the existing designs into IIoT with the help of Wireless Sensors and Actuator Networks. This research focuses on securing communication and confidence among the nodes for industrial applications [11]. It is aimed to improve issues like scalability, security, dependability, energy efficiency, and interoperability which are expected to collaborate with the industrial applications and will help in providing a secure and efficient means for communication. In this research, a trust management model is proposed over IIoT by using an energy-efficient access control scheme. Additionally, this research aims at improving the existent IoT model energy and delays constraints.

The rest of this paper is organized as follows: Section II covers the Internet of Things, security protocols are covered in Section III. Methods and materials are covered in Section IV. Section V describes the method used to manage trust and detect attacks in IIoT. Section VI covers the results and discussion. Conclusions and possibilities for future work are covered in Section VII.

II. Internet of Things

For years, the Internet of Things (IoT) has been the technology of interest for innovating numerous other technologies. Innovation in industrial applications was

emphasized to ensure that technological advancement has been facilitated with the latest technology. However, issues like energy efficiency, delays, security, and safe communication are the main hindrances that need to be improved to successfully implement Trust management in IIoT [12].

IoT is a vast topic that cannot be defined in a single definition. A possible definition of IoT is a collection of services that connect objects, whether electrical, electronic, or non-electrical, to offer contextual services and seamless communication. Development of services like mobile phones, actuators, sensors, and Radio Frequency Identification (RFID) tags can help assess IoT's facility to facilitate human needs. It is also known as a network of embedded sensors for increasing the ease of connectivity. IoT also means 'internalization of every connected object. It allows humans to control the means of communication and data transfer. Even if the objects are modified, or technology is enhanced, there are possible chances that human involvement will not be required.

In this research, the following definition of IoT will be adopted: A device connected to an IoT network is a resource-constrained embedded system with the ability to perform multiple tasks simultaneously like signal processing, networking, and sensing. Batteries usually power it to consume lesser power and provide wireless communication capabilities.

The concept of IoT may change or evolve due to changing software and hardware technologies or the need for industrial applications to evolve. Implementing IoT in industrial applications will help improve efficiency and offer trusted communication among devices as IoT faces issues like data hacking or external intrusion [13]. If this happens with any industrial application, there are chances that it might damage a device or make it vulnerable to data theft. This research investigates how secure and trusted communication will be implemented in IIoT.

The concept of IoT involves numerous software components; however, the latest evolution has been made in link layer and application layer protocols and operating systems. Recent advances in the IoT domain, such as the Application Protocols**,** the OSI model represents that the application layer is known to be the abstraction layer known for acting as an interface between the applications running on the host and how it is communicating with the user.

The following list includes known application layer protocols for IoT [14].

- RESTful HTTP is the first IoT protocol acknowledged for executing Hypertext Transfer Protocol (HTTP). It is mainly used for web-based services in which most of the work is to facilitate communication between the client and the user. The transport layer is deployed in the TCP protocol. However, the usage of XML makes it inefficient for low-power purposes and complex for general usage. The latest improvements made in HTTP have enabled the header compression to improve the overall performance of the HTTP protocol. The overall power consumption issue has been suitably dealt with, but it is still inefficient for implementation in a resource-constrained device like IoT [15, 16].
- MQTT: based on a client broker-server architecture, the MQ Telemetry Transport protocol created by IBM is implemented using two types of communication processes, i.e., Publish/Subscribe and in HTTP as Request/Response. This protocol still uses TCP, but it is more efficient than HTTP.
- Jabber: this protocol was developed by an open-source community to support instant messaging. Similar to MQTT, communication depends upon XML. It supports the client-server model using both communications mediums, i.e., Request/Response and Publish/Subscribe. However, this protocol also uses TCP in the transport layer [17].
- XMPP: Jabber protocol was modified by Internet Engineering Task Force (IETF) by including SASL for authentication and TLS for communication encryption. It is supported in extensible messaging and presence protocol.
- MQTT-SN: IBM proposed a modified UDP-based version of MQTT, which is more efficient and used in Sensor networks.
- Web-Sockets: This protocol was designed to improve communication between web servers and browsers; however, apart from these services, it can be used independently as a client-server application protocol. This protocol also relies on TCP for the Transport layer.
- CoAP: The Constrained Application Protocol (CoAP) was developed for optimizing the efficiency of communication in WSN. This protocol, known as the Restful-based protocol, has been enabled to execute its services directly on network nodes. Depending upon the client-server model, it observes methods, and depending upon these methods; it allows the Request/Response procedure. Unlike other protocols, this protocol uses UDP protocol instead of TCP protocol in the Transport layer [18].

Link-layer protocols: The innovation of wireless technologies like Bluetooth and Wi-Fi has introduced Wireless Local Area Networks (WLANs). It has served as an optimal technique for every mobile sensing platform and a gateway for both techniques. However, the only barrier faced using this technology is power consumption. The device's battery dies down within a lesser time when these techniques are not used. Thus, it lessens the time for consumption of Bluetooth and Wi-Fi in any device. Numerous new improvements have been made in hardware components to reduce power consumption, like the one manufactured by a Texas instrument named CC3000. It has a reception consumption of 331 mW and a transmission consumption of 936 mW. The wireless technology consumes the majority of the power of any IoT device; consequently, while selecting the wireless technology for the device, it shall be observed how it affects the power. The device can operate for an extended period without exhausting its batteries. The IETF in 2006 developed a link-layer protocol 6LoWPAN with header compression and encapsulation. The primary purpose behind it was to use IPV6

networks. It was the most significant innovation for creating IP wireless networks for low-power devices [19].

Since the past decade, the use of embedded systems in industrial applications and other evolving technologies like upgrading cell phones has emphasized the innovative development of these systems. It can be seen currently that smart homes, safe cities, and automated gates result from these embedded systems. For this purpose, the hardware has been improvised, which will be discussed in this section.

The innovation of microprocessors and microcontrollers proposed the latest smaller technique and used lesser computational power. Previously, the IoT devices have been using microcontrollers due to lower computation power; however, with the current need to implement IoT in industrial applications for better and safe means of communication, the need for Microprocessors was felt. Intel Atom and the ARM Cortex-M73 result from this innovation, which consumes less power and is high-performance. Microprocessors are recommended for nodes that require a high level of power in the processing and mitigate the overhead in the communication [3].

The CoAP services are located in the link-layer and are used to design web-based services capable of working with resource-constrained devices. This technique is efficient for microcontrollers that can run over 6LoWPAN network stacks and have a small ROM and RAM. However, it gives high error rates in a packet transfer. The devices using this technology can switch to sleep mode to save power and give optimal performance for low-power networking. The Request/Response interaction model is provided by CoAP between the communication ends of the applications. This protocol supports key Web concepts, built-in discovery, extensible header options, and RESTful interactions. For integration with the web, CoAP can easily develop an interface with HTTP; it will help fulfill the needs of constrained environments like very low overhead, multicast support, and simplicity of procedures. The features of CoAP which are relevant to the research are discussed below:

- Two types of messages are transmitted, a confirming message with the exponential expiry time to receive the acknowledged message. On the other hand, a validation message is sent without the expected response from the server.
- Uniform Resource Identifier (URI) format uses specialized service endpoints and standard services. One example explaining the procedure is /.well-known/core path in RFC 5785, and the name Core Format knows another format.
- CoAP can send large messages in blocks with the stop and wait for mechanism. In this way, no data packet will be lost, and the complete message will be transmitted in 'Block wise transfers' [4]. Fig. 2 illustrates the transfer of message block-wise.

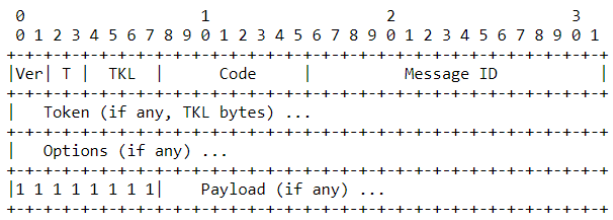


Fig. 2. CoAP Packet Format [3].

Providing E2E security is a widely discussed topic in conventional communication using the Internet. Though E2E was explored a long time ago, little research is available on E2E security using 6LoWPANS. The lossy nature of wireless links and the device's resource constraints are the main reasons for not applying E2E security mechanisms details to 6LoWPANs. IP-based IoT faces security challenges during the handshake process. To resolve security issues, it is suggested to: (1) validate certificates at the trusted 6BR, (2) a full handshake shall be avoided by session resumption, and (3) the owner of the resource-constrained device shall allow the handshake procedure. The certificate-based authentication is feasible for this type of authentication [4].

Due to heterogeneity in IoT, it becomes difficult to connect resource-constrained devices in a more secure and reliable way. Especially when it comes to the connection in industrial applications, apart from operating in a resource-constrained environment, the ruggedness of the environment and weather have to be considered. To enable this process to be implemented in industrial applications to ensure trusted communication between nodes, the IETF has proposed techniques using existing protocols like CoAP, the IPv6 Routing Protocol, and 6LoWPAN. These protocols proved to be useful for lossy and low-power networks. The Datagram Transport Layer Security (DTLS) protocol guarantees E2E security for different applications running on the same machine. It operates between the application layer and the transport layer. The DTLS consists of two layers; the upper layer includes any three stated protocols like application data, ChangeCipherSpec, Handshake, and alert.

The ChangeCipherSpec indicates that the Record protocol should protect the messages with security keys and a newly negotiated cipher suite during the handshake procedure. DTLS uses the alert protocol for communicating error messages due to lossy networks within the DTLS layers. Once the handshake procedure is completed, the record header is mainly responsible for cryptographically protecting the application data or upper-layer protocols. The record protection protocol offers authenticity, integrity protection, and confidentiality features. The handshake procedure is a chattier procedure than a DTLS protocol as it releases numerous messages in a synchronized fashion [4].

III. Security Protocols

A. Wireless Personal Area Networks 6LoWPAN

Fragmentation and header compression mechanisms of IPv6 datagrams are defined within the 6LoWPAN standard. The IPv6-connected WSNs are also known as IPV6 networks. The compression mechanism used in this protocol is Next

Header Compression and IP Header Compression (IPHC). Like IPHC, DTLS header compression can be applied between 6BR and sensor nodes within the 6LoWPAN networks. The complete information required for routing has been extracted from the IP layer. It happens because DTLS headers are part of the payload scheme. 6LoWPAN header compression mechanisms compress the headers in a UDP payload. To perform 6LoWPAN compression, a new modification is required in which a new NHC for UDP with different ID bits is assigned. This approach will extend the existing 6LoWPAN and be easier to implement than making changes to the existing technique [20]. Table I shows the reviewed protocols.

TABLE I. REVIEWED PROTOCOLS

Year	***Protocol***
1999	MQTT client broker-server architecture [21]
1999	Jeremie Miller announces the existence of Jabber [22]
2000	Roy Fielding first presented RESTful [21]
2004	Publishing XMPP standards: a modified version of Jabber protocol [23]
2011	WebSocket improved computer communications protocol [24]
2012	Trust-based communication – WSNs [25]
2012	DTLS a protocol that guarantees the implementation of E2E security [26]
2014	Constrained Application Protocol (CoAP) [27]
2017	6LoWPAN an approach for routing IPv6 over low-power wireless networks [28]
2017	Enhanced Lightweight DTLS for IoT [29, 30]

B. E-Lithe

IoT made the connection of millions of devices possible. However, developing secure communication is a challenge for IoT devices. If secure and trusted communication between Industrial applications using IoT is not possible, it does not only threaten productivity and efficiency, but it also threatens important data used within these industrial applications. It is proposed to provide secure communication within the IoT environment by implementing DTLS while constructing a secure transport layer over the datagram. DTLS is a protocol that is expected to provide secure communication in client-server applications. The mechanism depends upon transport layer security which prevents fragmentation, tampering, and message forgery. This protocol also deals with the datagram's size, loss of datagram, and packet re-ordering. However, an issue is identified that the DTLS protocol is defenseless against the Denial-of-Service (DoS) attacks [31] and requires more computation than an average device operation while working in a resource-constrained device. DoS attacks prevent the communication between two nodes and can disrupt the network services, thus, disrupting the communication between complete networks. DoS attack is identified when the requested services are not provided to the user due to an attack on one of the networking devices [8].

To overcome DTLS shortcomings for constrained devices, an Enhanced and lightweight DTLS protocol was proposed and named Enhanced Lightweight DTLS for IoT (E-Lithe). For this research, a trusted third part element will be added to E-Lithe for implementing E-Lithe in IIoT to manage trusted communication. The trusted third-party feature aimed to prevent the DoS attack by pre-sharing the secret keys. The E-Lithe protocol is explained as below:

- The server and the third-party trusted protocol agree on sharing a secret key before beginning the handshake procedure.
- A mutual secret key is shared between the client and a Trusted Third Party.
- The sharing of the mutual key prevents the power exhaustion of devices and authenticates the client-server communication.
- The client sends a handshake message to the server.
- If the server confirms the validity of the key, the server generates a hello message in return for the client response. However, if the keys are not matched, the process is terminated.

E-lithe uses lesser power during message transmission to prevent the overloading of fragmentation by applying the compression technique to ensure the lightweight transmission of messages. The compression strategy used for E-Lithe comprises a client Hello, a handshake layer, and a record layer. On the other hand, the record layer comprises a fragment, a sequence number, and an epoch. The handshake layer consists of message sequence and message type. The message is sent precisely with message type, and length details are ignored. Fig. 3 depicts the communication mechanism in E-Lithe.

C. Distributed Trust Management System

The devices connected through IoT face the issue of secure communication. Insecure communication could bring more devastating damage if it happens in industrial applications. It does not only damage the device, but it can also give access to sensitive data. The main issue is to identify malicious attacks before the handshake procedure. These malicious nodes choose selective attacks which require lesser processing requirements.

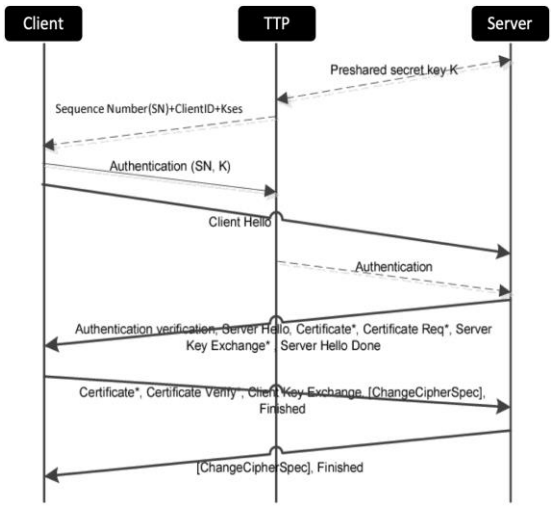


Fig. 3. Communication in the E-Lithe Scheme [8].

Trust management will check the fault in the network and focus on protecting nodes and networking connectivity. The main scheme for trust management is to implement trust among the connected devices. While checking the trust management, they also focus on checking the behavior of malicious nodes. In IoT, several trust management schemes are available, like centralized, decentralized, and hybrid.

The trust management value of each node is calculated on direct observation of the nearest nodes. This value is zero at the start. The start value shows no trust between any two devices at the beginning as trust needs to be built. An announcement is sent to the nearby nodes, and the value is calculated through the sending nodes. The service provided by nodes is denoted as a healthy node and is considered a broken node if the service is not provided on time.

A policy-based secure and trustworthy sensing scheme called "Real Alert" has been observed for this research. In the mechanism, IoT node attributes and the data trustworthiness are calculated through evaluating the anomalous data and the contextual information from which this data is collected. The monitored direct trust value measured from network communication relies on the quantitative value of the trust model. The evaluated features are integrity and delay, consistency of the packet content, repetition rate, and packet forwarding capacity. The D-S theory is computed to calculate trust. However, the drawback of this scheme is that it uses a large amount of data, and streaming this amount of data creates a problem for the conventional networking system.

The redemption scheme and the trust management differentiate between malicious behaviors to detect and defend against On and Off attacks and temporary errors. The ratio of good behavior to the total behavior is calculated using the difference as the predictability trust, and a static sliding window is used for recording previous behavior [32].

D. Naïve's Bayes Theorem

Machine learning algorithms are proposed to be an alternative for calculating trust values among the connected nodes. This theorem can help from attempting multiple calculations. The BAN-Trust scheme will have opted for this scheme on the recommendation of other nodes detecting an off behavior of any node. In this process, if any node identifies and reports that the other particular node is behaving off or sending flooding messages, the system can ban that node from the communication network. In this way, the rest of the network is protected [33, 34].

A Naïve's Bayes trust management model is easier to build and is less complex. It can use large datasets and does not rely on iterative parameter estimation. The Bayesian theorem is stated as below:

$$P(x|y) = \frac{P(x|y) \times P(x)}{P(y)} \quad (1)$$

Where $(x|y)$ = probability of a class, x given instance, y, $p(y|x)$= probability of instance, y given class and x, $p(x)$=probability of occurrence of class x, $p(y)$=probability of instance y occurring.

The Bayesian theorem uses one parameter only. This parameter calculates all features and simplifies them through the Naïve Bayes theorem for numerous features. This theorem can be used for detecting malicious nodes. The features of the Naïve Bayes classifier are Packet Loss Rate and the Packet Error rate. Malicious nodes intentionally disseminate erroneous packets or drop packets. For this purpose, the packet loss rate and the packet error rate are included for calculating the trust value between nodes. According to the Bayesian theorem, trustworthiness can be classified as High, Low, or Moderate [33]. The calculation of the Level of Trust in Naïve Bayes is explained in Fig. 4.

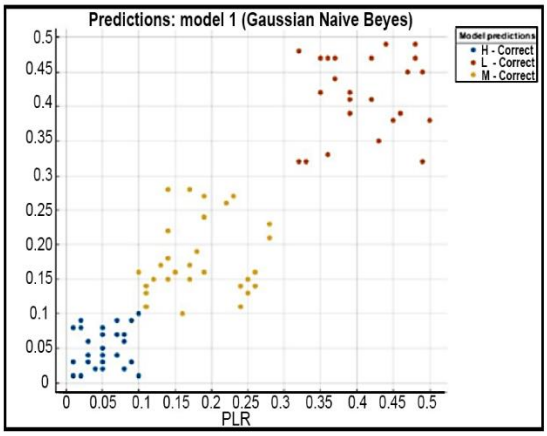


Fig. 4. Classification of Trust Values after Training the Model.

IV. Methods and Materials

To present the real-time implementation of the improvised E-Lithe, Contiki is used. It is an open-source operating system used for implementing IoT. The proposed header compression technique can be implemented with the support of 6LoWPAN. Cooja enables cross-level simulations at many levels and supports the 6LoWPAN protocol with a more convenient interface. Cooja can combine low-level and high-level simulations of sensor node hardware and how they behave in a single simulation. The Cooja simulator provided a flexible approach for implementing improved E-Lithe for trust management in IIoT. The partial real-time scenario generated with the help of Cooja will give a better insight for deploying this research at the industrial level. The improved lithe implementation requires the support of four components: CoAP, DTLS, DTLS header compression, and the CoAP-DTLS integration module. Open-source Ubuntu 14.04 LTS 64bit will be used for DTLS implementation. It uses the pre-shared keys: TLS_PSK_WITH_AES_128_CMC_8 for supporting the basic cipher suite. For the WiSMote Platform, Ubuntu 14.04 LTS 64bit and VM Ware workstation are used. The default CoAP implementation will be used for CoAP implementation in Conitki. An integration module will create a collaboration between DTLS and CoAP and enable the CoAP protocol. Independent application access is created due to this integration with CoAP. In this process, the CoAP messages are handed over to DTLS, responsible for transmitting them to the receiver's end [35].

Initially, all the CoAP messages are received at DTLS. Once processed and checked, DTLS transfers these messages to CoAP, stored at the application layer. The header compression will be used, as an extension, to implement 6LoWPAN in Contiki. The 6LoWPAN layer has been placed between the Medium Access Control and IP layers. The packets ready to be transmitted from the nodes in the IP layer are known as Output packets. The packet received at the node from the MAC layer is known as the input packets. However, the 6LoWPAN layer can process the UDP packets from both directions. The UDP packets depending upon the messages, are divided into two categories. The default DTLS port for pre-configured input packets identifies CoAP messages. In addition to it, the security shall not be compromised during the E2E sharing of keys during the header compression scheme. Yassine's is a secure version of DTLS [35]. The Yassine's version of DTLS and E-Lithe are observed to have closer values, as shown in Fig. 5. However, when compared and shown results through the graph depicted in Fig. 5, it has been identified that the E-Lithe can handle the compressed data packets better than the heavy data packets. E-Lithe uses the header compression scheme despite the cookie exchange scheme, which helps it perform better.

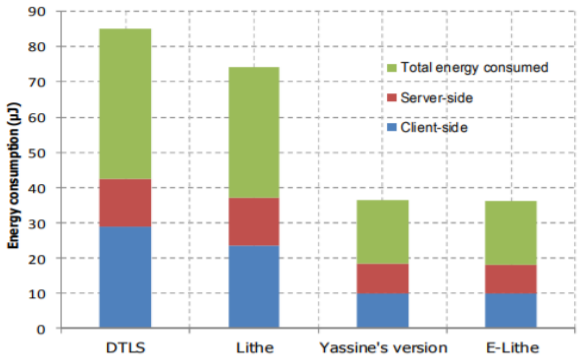


Fig. 5. Total Energy Consumed for Each DTLS Variant.

Improved Lithe will be evaluated by implementing sensor nodes in Contiki. As a hardware platform, WiSMote will be used. WiSMote proposes the features like a 16-bit RISC microcontroller, MSP430 5-Series, 16 MHz, an IEEE 802.15.4 (CC2520) transceiver, and 128/16 kB of ROM/RAM. WiSMote was selected due to the RAM and ROM requirements of DTLS.

The network will consist of forty WiSMote. One of the WiSMote will serve as a server that will communicate with other nodes. As this research aims to identify the broken nodes and the attacked nodes, these nodes will be sending *Hello Flood* Messages to other nodes. The attacked nodes are essential to be identified because they are the loop through which data theft can happen, or any foreign intrusion is expected. The communication between nodes that are attacked or broken will be restricted. They will not be able to send messages. The rest of the nodes performing on time can be called trusted nodes. It is essential to ensure that the nodes within the network are trusted because a single broken or attacked node within an industrial network can damage the complete network and may cost resources for replacing the hardware or cause downtime.

V. A Smart Trust Management Method to Detect On-Off Attacks in the IIoT Applications

The proposed approach in this research aims to detect on and off attacks and broken nodes on networks for IIoT. The communication between the industrial applications will be calculated through the available metadata attributes. IIoT metadata can be evaluated by sending it to the proposed algorithm.

Data is entered into the feature type extraction process for the pre-processing phase. Hashing vectorizer is used for processing text data. The text in this format is converted into token occurrences of a matrix. The integer index mapping string is named by the token string name. This approach is used because it does not need the support of the dictionary and can be used for streaming. The pre-processed dataset is fed to a machine learning classifier for identifying the class. Few limitations have been accepted for evaluating the industrial data, such as calculating the average temperature for a city. If the temperature is within range, it will be called trusted data, but if it is out of range, it can be labeled as a broken or attacked response, although it could result from extreme, unusual weather conditions. A decision function value is returned if a classifier confirms an identified class. The methods adopted the use of the decision function for calculating the size of the Elastic Slide Window. It is evaluated by observing the model decision function of the distance hyperplane of the sample data. A high positive decision value is received corresponds to high prediction assurance, as illustrated in Fig. 6 and Fig. 7 [4].

Fig. 6. Expected Range of Trusted Value [4].

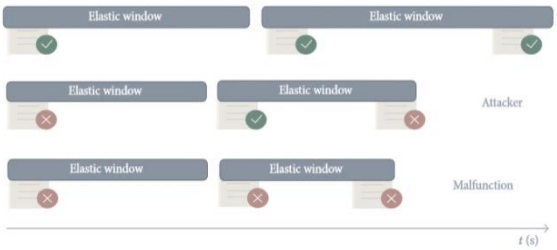


Fig. 7. The Elastic Slide Window [4].

The introduced Elastic Slide Window is an essential concept for data flow. Using the time frame analysis, it enhances trust. All the values are either good or bad during the On and Off attack are sent in a discretionary manner. If the system is healthy, it will accept the good value over time, but the suspect of being an attack is always expected. If the identifier sends a low decision function value or an identified class, the trust for that value is doubted, and that particular resource is either expected to be tested again for trust, or its communication is limited. The decision function values evaluate the size of the elastic slide windows. Any time a low

decision function value is received, the elastic slide window is increased for evaluating the trust of that particular decision function. The trust dispatcher records the storage size of the elastic slide window in the database. This feature also determines the trust response, i.e., Good, Broken, or on and off attacker. The proposed algorithm is illustrated in Fig. 8.

```
input: A metadata m (ID, read)
output: A predicted type: (Tusted, On-OffAttacker, Broken)
(1)  eswAlpha ← alpha;
(2)  eswInit ← beta;
(3)  NewPrediction ← Classifier.Predict(m);
(4)  NewDecisionFunction ← Classifier.DecisionFunction(m);
(5)  if m in Database then
(6)      m.SlideWindow ←
           (eswInit + time()) − NewDecisionFunction;
(7)      m.prediction ← NewPrediction;
(8)      if m.SlideWindow ≥ Time() then
(9)          if NewPrediction == −1 and m.prediction == −1 and
               NewDecisionFunction ≤ eswAlpha then
(10)             m.prediction ← 0;
(11)         end
(12)         if NewPrediction != m.prediction and
               NewDecisionFunction ≥ eswAlpha then
(13)             m.prediction ← −1;
(14)         end
(15)         if NewPrediction != m.prediction and
               NewDecisionFunction ≤ eswAlpha then
(16)             m.prediction ← m.prediction;
(17)         end
(18)     end
(19) end
(20) m.SlideWindow ← m.SlideWindow − NewDecisionFunction;
```

Fig. 8. The Smart Trust Management Algorithm for Industrial Internet of Things.

The smart trust management server will consult, via the Constrained Application Protocol, for an object, and the value is returned in JSON formatted data. Honesty, exploitation, and selfishness levels determine the node's trust. The node will be tagged as an attacker node if the trust value is not satisfied with the threshold value. For example, the object Id: 15 with a metadata payload of 45 degrees Celsius is marked as an On and Off attacker. The decision function for a trust score value is also presented. Other expected results can be transcribed as Good for predictable devices or even two broken nodes in the same elastic slide window.

VI. Results and Discussion

This research's proposed trust management method is expected to detect On and Off attacks in IIoT with 97.1% precision tested on a real-time dataset. For the simulated environment, 95% of precision is achieved. The reasons behind choosing the proposed method were: (1) to establish secure communication with resources limitation presence in IIoT, (2) to enhance the defenseless DTLS protocol against DoS attacks to gain the advantage of lightweight transmission with low power consumption. Compared to other studies, the proposed method is 95% faster, and On and Off attack is predicted 5% more accurate in On and Off identification attacks. The Elastic Slide Window feature helped identify the malfunctioning or broken nodes among the misbehaving devices by evaluating the exploitation, selfishness, and dishonesty levels. Fig. 9 and Table II show the simulated node behavior based on delay and overhead.

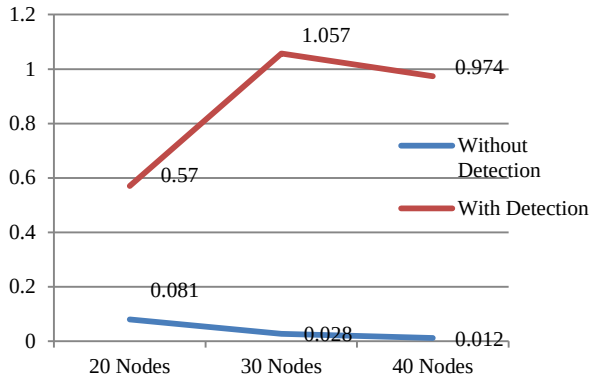


Fig. 9. Power Consumption in Nodes vs. Delay.

TABLE II. The Relationship between the Number of Nodes and Delay Time with and without Detection

Number of Nodes	*Without Detection*	*With Detection*
20	0.081	0.57
30	0.028	1.057
40	0.012	0.947

The decision function boundaries are created by joining the trained datasets, which are blue in the illustration. The x-axis in Fig. 9 represents the number of nodes, whereas y-axis represents delay time for communication. The OneClassSVM classifier efficiently grouped the test data, which was near to the trained dataset. The abnormal samples opted in this research were presented far from the decision function values. Typical or trusted values are near 0 for an identified class. The broken, attacked, or misbehaving nodes give high distance values up to -200 [3].

The proposed method could be affected by suspected validation threats like datasets, classifiers seeds, and random simulation outputs. The output of each simulation varies according to its run time. To minimize the error in the validation threat, the simulation is executed three times, and the average of these three simulations is used as a result. During the training tasks of the models, scikit-learn library procedures create an alert for the users that they use parameter initialization variable and random seed values, which together contribute towards different types of precision values in results. The average values were annotated for three fitting rounds for the simulation values. The null values were sanitized from the real-world dataset by removing the non-related data and NaN values. However, the related work sessions could be threatened if no relevant study is considered during the concept development. To minimize the risk, main indexing databases were considered for verifying references and citations of sources within the context to validate the theories.

The previous sections discuss the improvisation in the most well-known technique, 'Lithe,' to implement security and trust-based communication in IIoT. The selected trust management technique is analyzed and evaluated, relying on essential trust metrics: scalability, availability, adaptability, reliability, privacy, integrity, and accuracy. The selected articles show that the researchers have focused on implementing security and no

data loss communication between nodes. While researching the available data, it has been identified that there is limited data available in research regarding Trust management in IIoT. The security feature could be established for a smaller network. However, the industrial network is massive. It needs to be protected and ensured by implementing improved E-Lithe to prevent DoS attacks on devices and establish that the devices within the network could be trusted because they have been responding within range and within the expected time.

Table III illustrates that the researchers have focused on a few parameters for conducting research like scalability, adaptability, availability, accuracy, and security. However, the features like establishing the fact that the nodes in the network are not broken, attacked, or misbehaving nodes, which can be dangerous for the complete network and its authenticity, are not discussed. The attacks not only threaten the safety and secure transmission of sensitive data but also damage the network device.

TABLE III. PARAMETERS FOR SIMULATION

Network Simulator Parameters		
Parameter	***Value/Description***	***Remarks***
Number of nodes	40 nodes	255 for each scheme
Simulation area	1000 x 1000 m	Controlled by wireless coverage
Topology	Random	Determined by Cooja (Simulator)
Radio medium	UDGM	Directed Graph Radio Medium
Routing protocol	RPL	IPv6 Routing Protocol for Low power and Lossy Networks
Mote type	WiSMote	Contiki-wismote-platform
Packet analyzer	Wireshark	Network protocol analyzer
Packet interval	10 seconds	10 ms to 60 S

Due to the fast growth of the technology and the network requirements, the IoT network cannot handle the trustworthiness computation. The implementation of Lithe ensured the secure communication of data, but this technique does not support the trusted communication of data between devices. The technique faces the issues like fake recommendations from other devices. A hacked device can send messages for other devices that they are hacked, limit their communication, downgrade the ranking, which may not be part of the system, or upgrade the rank of any device once they are fit for use. Due to this, they may predict the wrong trust value.

The trust decision is developed on static rules, whereas the trust decision must be taken dynamically during trust negotiations. Due to its constrained environment, the improved E-lithe cannot be directly applied in the IoT. The devices need to be upgraded at the application and the transport layers to prevent loss in the data packets. Even if the message is paused, it does not take much time. The message reaches its destination in a complete form.

In this section, Cooja's implementation will be discussed. Initially, the installation of the software will be explained. For implementing the trusted communication between IIoT, Contiki 3.0 open-source OS is used with supporting software of VMware. The aim was to generate simulation with the pre-processed data for identifying how nodes can communicate by managing the trust factor. It is a partial real time scenario. All the variations will be tested here so that once the program is fully completed, it is launched easily for the real-time industry without the fear of failing. Studying traffic volume, position, and number of malicious nodes at least requires 120 minutes for determining the On and Off attacks. However, the method proposed is expected to detect the On and Off attacks in approximately 10 minutes which is 95% faster than other methods. The method offers a faster way for prediction and is 96% more accurate. Node31* (Good), 8* and 32* are (attackers); Nodes 5, 9, 22 are (Good) and 12, 13 and 15 are on and off attacked, therefore they cannot be trusted. It is generated from the simulated scenario as shown in Fig. 10.

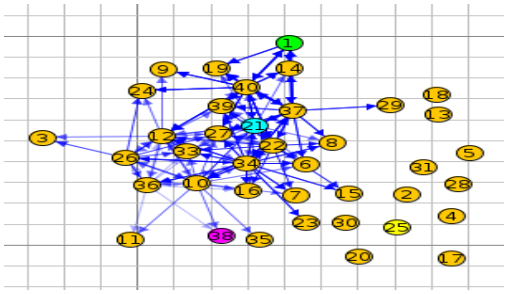


Fig. 10. DODAG Attacks Generated in Cooja Contiki 3.0.

Node*5 is identified as a selfish node and was identified in the first seven minutes of simulation execution. The positive trust score is related to good nodes, which could be trusted, and negative values are related to attacked nodes.

The above scenario is one of the probable situations in which not all the nodes are attacking nodes; they may be broken or selfish, and due to this, they have been unable to send messages within range. However, and to the best of our knowledge, there is no relevant research available regarding the message sent in reply that the particular node can be trusted or not and for broken nodes as well that the following node is broken and can participate. Thus, the trust-based mechanism for IIoT has been calculated depending on factors like Dishonesty, Exploitation, and Selfishness levels.

The dishonestly level depends upon the packet-dropping nature. If there is any packet loss during the communication process, the message will not be received in a complete form at the destination. It is pre-conditioned here that if the dishonesty level is greater than one, the node is suspected of being dishonest.

The exploitation level is calculated based on the over flooding of hello packets or Distributed Interactive Simulation (DIS) message. If Dishonesty Hack (EH) level is greater than 1, it is determined as a suspect.

Selfishness level is determined based on the attractive nature of the node. Typically root node alone had higher quality than other nodes. However, the attacker node proposed

a lower rank than the root node to attract the attackers. If the Selfishness hack (SH) level is greater than 1, it is determined as a suspect. Trust calculation:

$$Trust\ (Ti) = (Weight \times \text{Hi}) + (Weight \times \text{Ei}) + (Weight \times \text{Si}) \quad (2)$$

Where *Weight* factor = 0.33, H_i is Dishonesty level, E_i is Exploitation level, and S_i is Selfishness level. If trust is greater than 1, the node is determined as an attacker.

For validation purposes, the annotated dataset was found useful. The method was compared to other machine learning algorithms like linear SVM, Neural Net, Naïve Bayes, and K Neighbors Classifier. Comparisons are illustrated in Table IV.

The proposed method was able to identify two good nodes, two broken nodes, and three attacking nodes. Fig. 11 shows the types of nodes identified. The actual nodes are shown in the form of filled nodes. The color of nodes is depicted with names at the end of the image. The mark around the nodes presents the predicted class.

For developing and evaluating intelligent data middleware, we used 3111 samples of temperature data collected by 116 sensors from February to March 2019 for Arahnus in Denmark. The average temperature in Arahnus ranges from -3 to 16 degrees Celsius. A total of 501 misbehave samples were simulated using random out-of-range temperature observations. Fig. 12 presents the attacking dataset captured by Wireshark. It gave similar results to the simulation, and approximately 97% of precise results were generated.

TABLE IV. COMPARISON WITH SUPERVISED CLASSIFIERS

	Classifier				
	Linear SVM	***Naïve Bayes***	***Neural Net***	***Nearest Neighbors***	***Our Method***
Precision	0.88	0.92	0.92	0.91	0.96
Pecall	0.71	0.82	0.81	0.84	0.85
F1-Score	0.74	0.85	0.84	0.84	0.87

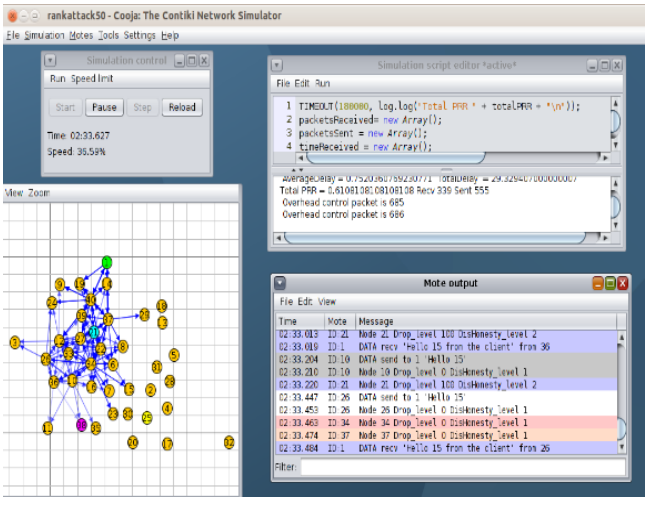


Fig. 11. Rank Attack for Nodes Cooja Contiki 3.0.

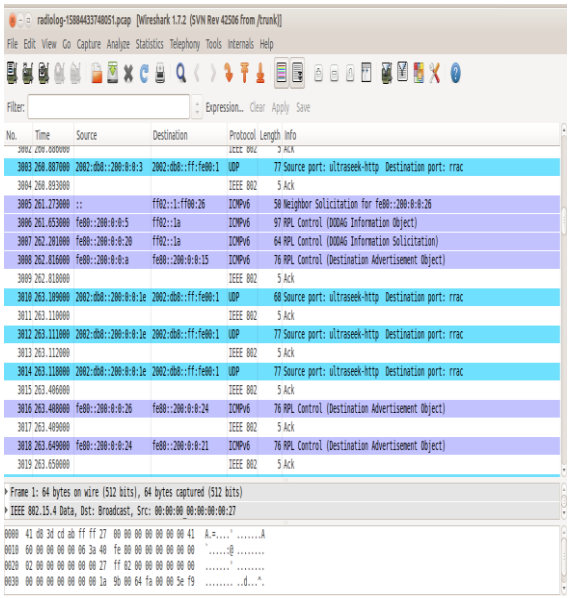


Fig. 12. Attacking Dataset Captured in Wireshark.

The output results from one-class classifiers types have been used in the proposed method. The OneClassSVMclassifier has been identified to help identify the attacking nodes and prove trusted communication between nodes. The one-class classifier is used to justify the aim that the trusted communication between the nodes has to be established by separating good nodes from the broken nodes and limiting the communication of malfunctioning nodes. However, it is hard to tag a classifier for each type of metadata.

VII. CONCLUSION AND FUTURE WORK

Implementing trust management in the Industrial Internet of Things (IIoT) is challenging. The relevant research available has been focused on implementing IIoT networks, but little focus has been made on IIoT trust among the nodes. Though this is a new concept, it is limited by the little data available. Using the current research available, a smart trust management system has been introduced using machine learning algorithms and size of Elastic slide window. The proposed method successfully detected On and Off attacks on nodes, malicious nodes, healthy nodes and broken nodes. It is necessary to identify the type of node within the network for identifying that if particular nodes could be trusted or not. The proposed technique was able to predict 97% of accurate behavior of nodes faster than other machine learning algorithms. The Elastic Slide window introduced has the capacity of identifying broken nodes, malfunctioning nodes or attacking nodes in industrial network as they have the capacity of handling bigger data and take the strain of massive communication.

This research examined the available research on network security, IoT technologies, Wireless Network Systems and authentication for IIoT, and the security protocols for lightweight key management through the preliminary study and literature survey. Then we proposed E2E key management using a lightweight security solution to achieve device authentication and communication in secure groups:

simulations, estimations, and real-time implementations evaluated and validated such solutions.

The proposed system can be tested against other machine learning techniques and scenarios in more complex industrial networks. Currently, only the trust factor is evaluated. If further research is conducted, there are possible chances of implementing an intranet network facility using E-Lithe approach for ensuring that the particular network can only be accessed by authorized people and can deal with any incoming attacks. Also a random dataset can be used to test if the mechanism is able to identify the validity and trust factor within the system. Also, elastic slide window can be used for identifying other IoT related trust based attacks in industrial environments like ballot-stuffing attacks, opportunistic service attacks, bad-mouthing attack and self-promotion attacks.

ACKNOWLEDGMENT

Thanks to the Saudi Electronic University for sponsoring this work.

REFERENCES

[1] S. Al-Rubaye, E. Kadhum, Q. Ni, and Anpalagan, A. "Industrial Internet of Things Driven by SDN Platform for Smart Grid Resiliency." *IEEE Internet Of Things Journal*, *6*(1), pp. 267-277, 2019. DOI: 10.1109/jiot.2017.2734903.

[2] G. Falco, C. Caldera, and H. Shrobe. "IIoT Cybersecurity Risk Modeling for SCADA Systems." *IEEE Internet Of Things Journal*, *5*(6), pp. 4486-4495, 2018. DOI: 10.1109/jiot.2018.2822842.

[3] M. Hasan, and H. Al-Rizzo. "Optimization of Sensor Deployment for Industrial Internet of Things Using a Multiswarm Algorithm." *IEEE Internet Of Things Journal*, *6*(6), pp. 10344-10362, 2019. DOI: 10.1109/jiot.2019.2938486.

[4] F. Liang, W. Yu, X. Liu, D. Griffith, and N. Golmie. "Towards Edge-Based Deep Learning in Industrial Internet of Things." *IEEE Internet Of Things Journal*, pp. 1-1, 2020. DOI: 10.1109/jiot.2019.2963635.

[5] X. Liu, H. Huang, F. Xiao, and Z. Ma. "A blockchain-based trust management with conditional privacy-preserving announcement scheme for VANETs." *IEEE Internet Of Things Journal*, pp. 1-1, 2019. DOI: 10.1109/jiot.2019.2957421.

[6] A. Alyousef, K. Srinivasan, M. S. Alrahhal, M. Alshammari, and M. Al-Akhras, "Preserving Location Privacy in the IoT against Advanced Attacks using Deep Learning" International Journal of Advanced Computer Science and Applications (IJACSA), 13(1), 2022.

[7] http://dx.doi.org/10.14569/IJACSA.2022.0130152

[8] P. Ray, M. Mukherjee, and L. Shu. "Internet of Things for Disaster Management: State-of-the-Art and Prospects." *IEEE Access*, *5*, pp. 18818-18835, 2017. DOI: 10.1109/access.2017.2752174.

[9] J. Mcginthy, and A. Michaels. "Secure Industrial Internet of Things Critical Infrastructure Node Design." *IEEE Internet Of Things Journal*, *6*(5), pp. 8021-8037, 2019. DOI: 10.1109/jiot.2019.2903242.

[10] M. Zhaofeng, W. Lingyun, W. Xiaochang, W. Zhen, and Z. Weizhe. "Blockchain-Enabled Decentralized Trust Management and Secure Usage Control of IoT Big Data." *IEEE Internet Of Things Journal*, pp. 1-1, 2020. DOI: 10.1109/jiot.2019.2960526.

[11] C. Zhu, J. Rodrigues, V. Leung, L. Shu, and L. Yang. "Trust-Based Communication for the Industrial Internet of Things." *IEEE Communications Magazine*, *56*(2), pp. 16-22, 2018. DOI: 10.1109/mcom.2018.1700592.

[12] H. Tschofenig, and E. Baccelli. "Cyberphysical Security for the Masses: A Survey of the Internet Protocol Suite for Internet of Things Security." *IEEE Security & Privacy*, *17*(5), pp. 47-57, 2019. DOI: 10.1109/msec.2019.2923973.

[13] P. K. Malik *et al.*, "Industrial Internet of Things in Industrial Revolution 4 . 0 : A State-of-The art in Review," *Computer Communications*, vol. 166, no. March 2021, pp. 125–139, 2019, DOI: 10.1016/j.comcom.2020.11.016.

[14] M. Alsahli, M. Almasri, M. Al-Akhras, A. Al-Issa, and M. Alawairdhi, "Evaluation of Machine Learning Algorithms for Intrusion Detection System in WSN," International Journal of Advanced Computer Science and Applications (IJACSA), 12(5), 2021.

[15] http://dx.doi.org/10.14569/IJACSA.2021.0120574.

[16] G. Fortino, M. Hassan, M. Zhou, A. Goscinski, M. Bhuiyan, J. Li, and S. Bhattacharya. "Guest Editorial Special Issue on Emerging Social Internet of Things: Recent Advances and Applications." *IEEE Internet of Things Journal*, *5*(4), pp. 2478-2482, 2018. DOI: 10.1109/jiot.2018.2860339.

[17] A. Praseed and P. S. Thilagam, "Multiplexed Asymmetric Attacks : Next-Generation DDoS on HTTP / 2 Servers," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1790–1800, 2019, DOI: 10.1109/TIFS.2019.2950121.

[18] F. Alsattam, M. Al-Akhras, M. Almasri, and M. Alawairdhi, "Rule-Based Approach to Detect IoT Malicious Files," Journal of Computer Science, 16(9), 2020.

[19] A. R. Calibration, D. Using, and E. Intelligence, "A Remote Calibration Device Using Edge Intelligence," *Sensors*, vol. 22, no. 1, pp. 1–17, 2022.

[20] F. Seidel and C. Meinel, "Deep En-Route Filtering of Constrained Application Protocol (CoAP) Messages on 6LoWPAN Border Routers," in *IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019, pp. 201–206.

[21] M. Gidlund, G. Hancke, M. Eldefrawy, and J. Akerberg. "Guest Editorial: Security, Privacy, and Trust for Industrial Internet of Things." *IEEE Transactions On Industrial Informatics*, *16*(1), pp. 625-628, 2020. DOI: 10.1109/tii.2019.2953241.

[22] D. Palma, "Enabling the Maritime Internet of Things : CoAP and 6LoWPAN Performance Over VHF Links," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 5205–5212, 2018, DOI: 10.1109/JIOT.2018.2868439.

[23] *MQTT Version 3.1.1 Plus Errata 01*. Edited by Andrew Banks and Rahul Gupta. 10 December 2015. OASIS Standard Incorporating Approved Errata 01. http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/errata01/os/mqtt-v3.1.1-errata01-os-complete.html.

[24] A. Oram, "Peer-to-peer: Harnessing the Benefits of a Disruptive Technology," O'Reilly, 2001.

[25] P. Millard, *XMPP Protocol XEP-0060: Publish-Subscribe.* XMPP Standards Foundation 2004.

[26] I. Fette, *The WebSocket Protocol.* Hampton, UK: Google, Inc, 2011.

[27] A. Bairagi and D. Chakroborti, "Trust based D2D communications for accessing services in Internet of Things," *2015 18th International Conference on Computer and Information Technology (ICCIT)*, 2015, pp. 50-54, DOI: 10.1109/ICCITechn.2015.7488041.

[28] E. Rescorla, "*Datagram Transport Layer Security Version 1.2.* Palo Alto, CA: RTFM, Inc.

[29] K. Shelby, "The Constrained Application Protocol (CoAP)," CA, USA: Internet Engineering Task Force (IETF), 2014.

[30] R. Thubert, "IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Paging Dispatch," Sophia Antipolis, France: Internet Engineering Task Force (IETF), 2016.

[31] A. Haroon, S. Akram, M. A. Shah, and A. Wahid, "E-Lithe: A Lightweight Secure DTLS for IoT," *IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pp. 1-5, 2017, DOI: 1.1109/VTCFall.2017.8288362.

[32] A. Haroon, S. Akram, S. Ali., and A. Wahid. "E-Lithe: A Lightweight Secure DTLS for IoT." *IEEE*, *23*(123), pp. 5, 2017.

[33] I. Alnuman, and M. Al-Akhras, "Machine Learning DDoS detection for Generated Internet of Things Dataset (IoT Dat)," 2020 International Conference on Computer and Information Sciences (ICCIS), Jouf University, Jouf, 13-15 October 2020.

[34] F. Montori, L. Bedogni, and L. Bononi. "A Collaborative Internet of Things Architecture for Smart Cities and Environmental

Monitoring." *IEEE Internet Of Things Journal*, *5*(2), pp. 592-605, 2018. DOI: 10.1109/jiot.2017.2720855.

[35] B. Pourghebleh, K. Wakil, and N. Navimipou. "A Comprehensive Study on the Trust Management Techniques in the Internet of Things." *IEEE Internet of Things Journal*, *6*(6), pp. 9326-9337, 2019. DOI: 10.1109/jiot.2019.2933518.

[36] M. Al-Akhras, M. Alawairdhi, A. Alkoudari, and S. Atawneh, "Using Machine Learning to Build a Classification Model for IoT Networks to Detect Attack Signatures," International Journal of Computer Networks & Communications (IJCNC), 12(6), 2020.

[37] L. Chettri, and R. Bera. "A Comprehensive Survey on Internet of Things (IoT) Toward 5G Wireless Systems". *IEEE Internet Of Things Journal*, *7*(1), pp. 16-32, 2020. DOI: 10.1109/jiot.2019.2948888.

Machine Learning Application for Predicting Heart Attacks in Patients from Europe

Enrique Arturo Elescano-Avendaño[1], Freddy Edson Huamán-Leon[2], Gilson Andreson Vasquez-Torres[3]
Dayana Ysla-Espinoza[4], Enrique Lee Huamaní[5], Alexi Delgado[6]
Systems Engineer Program, Universidad de Ciencias y Humanidades, Lima-Perú[1, 2, 3, 4]
Image Processing Research Laboratory, Universidad de Ciencias y Humanidades, Lima Perú[5]
Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima-Perú[6]

***Abstract*—Even today, there are still a large number of people suffering from heart attacks, which have already claimed numerous lives worldwide. To examine the main components of this problem in an objective and timely manner, we chose to work with a methodology that relies on taking and learning from real and existing data for use in training and testing predictive models. This was carried out to obtain useful data for the present research work. There are in parallel different methodologies that do not quite fit the model of this work. Data was collected from the "Center for Machine Learning and Intelligent Systems" which in turn contains data from patients who have ever suffered a cardiovascular attack and from patients who never suffered the disease, all of them being patients selected from different medical institutions. With the corresponding information, it was subjected to different processes such as cleaning, preparation, and training with the data, to obtain a logistic regression type automatic learning model ready to predict whether or not a person may suffer a cardiovascular attack. Finally, a result of 87% accuracy was obtained for people who suffered a heart attack and an accuracy of 81% for people who would not suffer from this disease. This can greatly reduce the mortality rate due to infarction, by knowing the condition of a person who is unaware of his or her health situation and thus being able to take appropriate measures.**

***Keywords*—Prediction; machine learning model; logistic regression; heart attack**

I. Introduction

Nowadays it is more common to talk about people prone to cardiac arrest [1], the lifestyle of the general population has changed so drastically that people have started to develop cardiovascular problems frequently [2]. There are several factors to consider, one of the most obvious of which is the type of food that people choose to eat, such as junk food [3].

To analyze the main factors of this problem in an objective and timely manner, we chose to work with a meta-analysis methodology [4], This consists of taking and studying existing test data and sorting them to obtain data beneficial to our research [5]. Different research methodologies are not completely adapted to the model of our research, such as Design Thinking Methodology, which is a trial and error model, or the Ethnographic method, which is a controlled study of a sample of the population. That is why the meta-analysis methodology is ideal for the objective of our research [6].

As the main study sample in this work, we took data from various medical institutions in Europe to compare and analyze why and how cardiovascular diseases have been growing. We took data from the "Center for Machine Learning and Intelligent Systems" and acquired a CSV with the corresponding information [7], by doing this, we were able to structure the information to obtain statistical tables that help to understand the problem [8].

The main objective is to help prevent and study heart attacks in vulnerable patients in depth to reduce the mortality rate due to these diseases through concrete statistics that were implemented using machine learning.

The structure of the article is as follows: in Section II we will see the methodology, in Section III we will see the detailed case study through statistical tables, in Section IV we will present the conclusions and recommendations of the research and, finally, in Section V we have the references.

II. Methodology

A. Information Gathering

The first thing that was done to make the investigation of Heart Attacks and have a solution, was to obtain as much information on the subject, being these real cases where people were affected.

It should be noted that the information obtained is not data, since it still has to go through a severe filtering process, and using parameters, we will get the data already separated and grouped as appropriate [9].

The sources from which the information was acquired must be reliable, we cannot resort to any page of dubious information, since this can be detrimental to the investigation. [9]. We need truthful data that does not corrupt the real and specific objective we have.

B. Parameter Configuration

In this stage, we will import the libraries for the training of our data, which will be divided into two processes [10].

1) Input data: A collection of records containing features important to the Heart Attack problem, this data will be used during training to set up the model to make accurate predictions about new instances of similar data, the values in the input data are a direct part of the model [11].

2) Parameters: These are the variables that the selected machine learning technique uses to fit the data [10]. The parameters and model are optimized and tuned through the training process, run data, evaluate the accuracy and adapt until the best values are found.

3) Validation data: this model it provides us to keep the data, as test training, it will also be trained with the missing data, adjusting the validation data to finalize it will be evaluated according to the percentage acquired with the test data [12]. The data model is divided into three parts, the information will be prepared without nulls and gaps. The small volume of data will not be efficient for training.

C. Data Preparation

In this stage, we will extract data that will be important for the quality of our result, so we will obtain better results and a high range of prediction positions [13].

This section is where the data was obtained or collected from different sources, such as databases, blogs, websites, spreadsheets, etc. [14]. To clean the data obtained and later have as a result a file with a format that we have applied as CSV.

D. Model Approach

The problem shows the frequency of cardiac problems in different groups of people, being the main problem the heart attack, for this problem we propose the decision making of the machine learning methodology logistic regression.

The logistic regression model is used for classification, it is a supervised type algorithm. This model is used when our objective is to forecast the probability of a certain event occurring or not [15].

This will help us to classify the data and perform automatic learning, being supervised. With this logistic regression model, we predicted the heart attack patterns, using logistic regression, the following are performed.

III. Case Study

A. Information Gathering

A dataset of available heart disease data from the following was used in this process.

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, MD
- Hospital Universitario, Zúrich, Suiza: William Steinbrunn, MD
- Hospital Universitario, Basilea, Suiza: Dr. Matthias Pfisterer
- Centro Médico VA, Long Beach y Cleveland Clinic Fundación: Robert Detrano, MD, Doctor [15].

In the aforementioned medical institutions, it was collected from different databases containing 76 attributes.

In particular, the database of the Cleveland institution was used to carry out our research; information on heart disease in patients. He concentrated on simply trying to distinguish the presence of heart disease [16].

B. Parameter Configuration

In this stage, the libraries to be used in the model were defined, as displayed in Fig. 1.

- Numpy: Provides functions for vector and matrix creation, especially mathematical operations.
- Pandas: Data handling, manipulation, and analysis.
- Matplotlib: Library for chart creation and data visualization.
- Scikit learn Library that will give us support for the creation and training of the machine learning model.
- Seaborn: It is a matplotlib-based library for the creation of graphs that provide a simple interface.

```
[1] import numpy as np # linear algebra
    import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
    import matplotlib.pyplot as plt
    import seaborn as sns
    from sklearn.preprocessing import LabelEncoder,StandardScaler
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import classification_report
    from sklearn.metrics import mean_absolute_error
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import confusion_matrix
    from sklearn.metrics import plot_roc_curve
```

Fig. 1. Import of Libraries.

C. Parameter Preparation

At this stage, the data from the institutions indicated in the following point were used (A); For this purpose, the CSV file was imported to our working directory, in this case, it will be saved in Google Drive, This will facilitate access to our data when running our predictive model. Using the Google Colab platform, which is a virtual machine environment based on Jupiter and Notebooks. This runs in the cloud, where we do the Python coding, as shown in Fig. 2.

```
[7] from google.colab import drive
    drive.mount('/content/drive')
    df=pd.read_csv(r'/content/drive/MyDrive/DataFrames/heart.csv')
```

Fig. 2. Reading the CSV File.

The following function was used pd.head()of the pandas bookstore to showcase the first 5 rows of the dataframe as displayed in Fig. 3.

```
[9] df.head()
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 3. Data Visualization.

As can be seen in Fig. 3, our dataset is displayed with a header divided into columns, which helps us to know what the values found in that column mean. In Table I, the meaning of the columns of the dataset is shown in more detail.

The next step is to perform data cleaning and data preparation using the function df.isna() that provides us with pandas to detect missing values, which will return values of type Boolean, which indicates missing or lost values, and the function sum()will return the sum of all these values, which will result in 0 if there are no missing values, as can be seen in Fig. 4.

TABLE I. UNDERSTANDING THE DATA

Description of the Data	
age	Age of patient
sex	Sex of the patient
exang	Exercise induced angina (1 = sí; 0 = no)
ca	number of important vessels (0-3)
cp	Type of chest pain 1.Typical anginaAngina atípica 2.No angina or angina 3. Asymptomatic
trtbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol mg/dl fetched via BMI
fbs	(Fasting Blood Suga >120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiography results Value 0: normal Value 1: tener ST-T saturation anomaly (T and wave versions /or ST elevation or depression of >0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Max. heart frequency
target	0= less chances of heart attack 1= more chances of heart attack

```
# help(df.isna().sum())
df.isna().sum()
```

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

Fig. 4. Missing Value Detection.

The next step is to verify that no duplicate values are found in the dataframe with the function df. duplicate(), which will return a result of Boolean type, which will tell us if there is the duplicity of data, and with the function sum(), will give us the sum of how many rows are duplicated, this case it turned out that we have a duplicate row, as seen in Fig. 5.

```
[10] df.duplicated().sum()

1
```

Fig. 5. Check for Duplicate Values.

The next step is to remove duplicate rows from the dataframe with the function df.drop_duplicates(inplace=True), as it visualizes the Fig. 6 duplicate data were deleted. Then we will check again if there are duplicate rows with the previous function df.duplicated().sum().

```
df.drop_duplicates(inplace=True)
print(df.duplicated().sum())

0
```

Fig. 6. Elimination of Duplicate Fields.

Fig. 7 shows the degree of a heart attack in older people who have higher blood pressure, higher cholesterol levels, lower maximum heart rate, under a thallium stress test, one way to quantify the degree of risk is to measure the discrepancy between the disease and non-disease distributions, based on logistic regression theory.

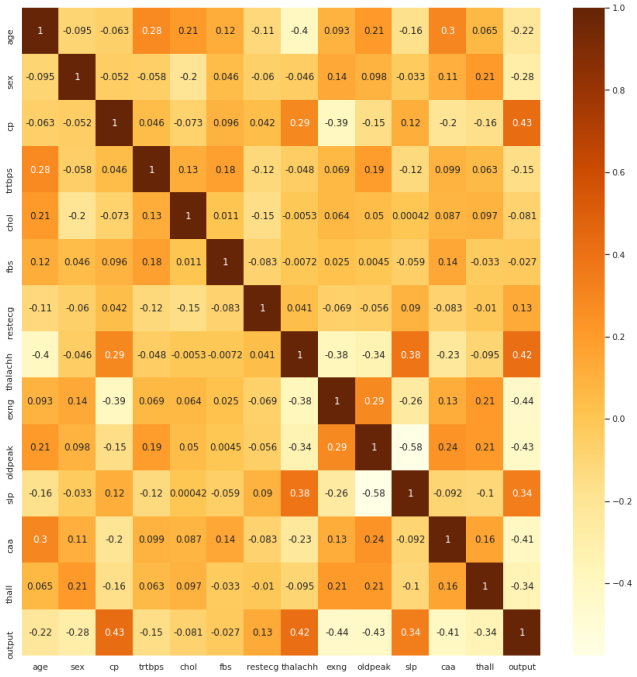


Fig. 7. Description of Data.

The creation of four graphs helped us to examine the most important characteristics that can be generated before a heart attack.

Using the matplotlib library, in Fig. 8 lines of code will be displayed, which will show us four plots shown in Fig. 9 from which useful information will be collected between "slp" - "output", "thalachh" - " output", "cp" - “output” and “old peak” - “output”.

```
fig,axes=plt.subplots(2,2,figsize=(20,20))

sns.kdeplot(ax=axes[0,0],x='slp',hue='output',data=df)
widths=[2,2]
g=sns.barplot(ax=axes[0,1],y='thalachh',x='output',hue='output',data=df)
g.legend(loc='center')

sns.countplot(ax=axes[1,0],x='cp',hue='output',data=df)

sns.swarmplot(ax=axes[1,1],y='oldpeak',x='output',hue='output',data=df)
```

Fig. 8. Graphics Creation Code.

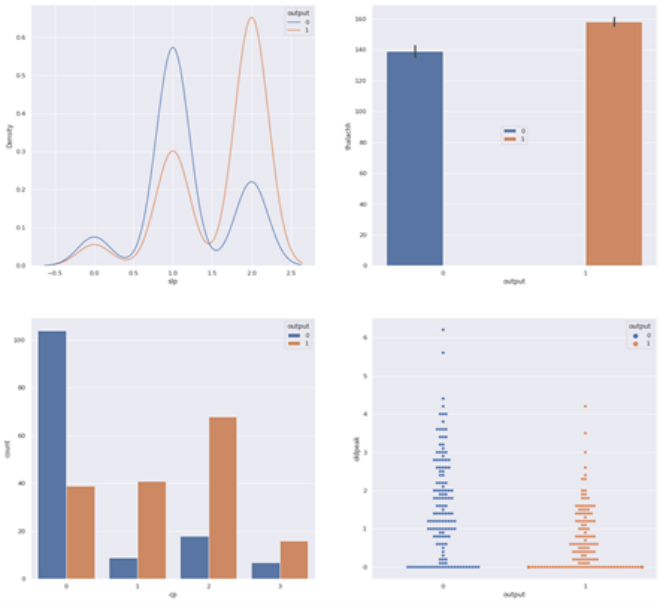


Fig. 9. Cardiac Symptom Score Charts.

The graphs indicate that:

- Patients more likely to have heart attacks tend to have higher heart rates.
- Patients with non-anginal chest pains are more likely to have a heart attack.
- The old speech distribution for both patient probabilities complements each other.

D. Units Approach to the Model

As shown in Fig.10 we carry out the preparation of the data, between training and testing, we create four variables: x_train de training, x_test of the test, and gives us a function train_test_split(), declarations 2 variables feature y output, we add the random percentage in the variable randon_state() el 2 percent to be applied to the output division. Now we define the training set with the function fit(x_train, y_train), we instantiate the logistic regression on a variable in this case named ClassiFier; now predict and with pred training and prediction, with which we obtained the report that shows the Fig.11, in which the summary of accuracy is displayed: recall, f1-score, support to see if you have the symptoms of heart attack, in this case, 0 means that you are not likely to have a heart attack. and 1 that if you are likely to have a heart attack.

As shown in Fig. 12 a function was created and inside we will perform the prediction and preparation with the new data. We make the confusion matrix for y_test and y_pred, it takes the index values and the columns of the data from the confusion matrix and we will put it in a graph for a better appreciation.

Use SI (MKS) or CGS as primary units. (SI units are recommended) English units can be used as secondary (in parentheses). An exception could be the use of English drives as a commercial identifier, such as a "3.5-inch disk."

```
[ ]  # Logistic Regression

     Scaleme= StandardScaler()
     features=df.drop(columns='output')
     output=df['output']

     X_train, X_test, y_train, y_test = train_test_split\
     (features, output, test_size = 0.2, random_state = 42)

     X_train=Scaleme.fit_transform(X_train)
     X_test=Scaleme.transform(X_test)

     Classifier=LogisticRegression(random_state=45)
     model=Classifier.fit(X_train,y_train)
     y_pred=Classifier.predict(X_test)
     print(classification_report(y_test,y_pred))
```

Fig. 10. Model Training.

```
              precision    recall  f1-score   support

           0       0.81      0.86      0.83        29
           1       0.87      0.81      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

Fig. 11. Model Training Report.

```
def cmcrcheck(X_test,y_test,y_pred,model):
    print(classification_report(y_test,y_pred))
    cm= confusion_matrix(y_test,y_pred)
    cmdf=pd.DataFrame(index=[0,1],columns=[0,1],data=cm)
    fig,axes=plt.subplots(figsize=(5,5))
    g=sns.heatmap(cmdf,annot=True,cmap='Greens',fmt='.0f',ax=axes,cbar=False)
    g.set_xlabel('Predicted Value')
    g.set_ylabel('True Value')

    plot_roc_curve(model,X_test,y_test)
    plt.show()


cmcrcheck(X_test,y_test,y_pred,model)
```

Fig. 12. Model Training Code.

Avoid combining SI and CGS units, such as current in Amps and magnetic field in Oersted. This often leads to confusion because the equation is not balanced in its magnitudes. If you must use mixed units, clearly state the units for each quantity you use in an equation.

Next, it is visualized in Fig. 13 that for 0 there was an accurate prediction of 83% and for 1 it had 84%. In our confusion matrix, he made 25 true positives, 26 true negatives, 6 false positives, and 4 false negatives.

	precision	recall	f1-score	support
0	0.81	0.86	0.83	29
1	0.87	0.81	0.84	32
accuracy			0.84	61
macro avg	0.84	0.84	0.84	61
weighted avg	0.84	0.84	0.84	61

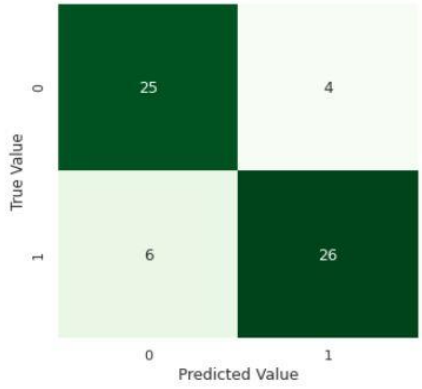


Fig. 13. Result of the Model.

IV. Case Study

A. From the Case Study Case

To see the precise dimensions of the research, it was compared with two other works, the first was with a Hybrid machine learning system [16] and the second with a Metaphorical machine learning system [17].

Fig. 14 shows the percentage of precision that was obtained in the results when applying the machine learning methodology, the blue color reflects the percentage of our research, the orange color is the percentage of the hybrid research, and the gray shows the work metaphorical.

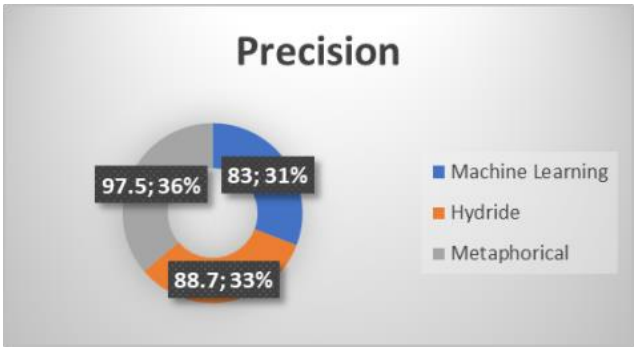


Fig. 14. Comparison of the Level of Precision with Other Works.

The value that was taken from each investigation was the level of precision of the analysis shown in percentages, resulting in similar comparisons, in the case of the metaphorical system, it uses simpler data, so its measurement is faster to carry out. The Hybrid system is developed with the union of different processes that result in a more complete analysis, and the work that was carried out is a direct machine learning implementation, so our data turns out to be more reliable and truthful, compared to the two other investigations.

B. Of the Methodology

The method used Machine learning is based on learning automatically, it provides us with tools that will help us make decisions according to the case analyzed, the logistic regression model is used, where the data is collected, after being analyzed, the configurations, data preparation and finally the problem statement.

This type of methodology used in all its phases has advantages and disadvantages, in Table II they are shown in better detail.

TABLE II. Advantages and Disadvantages

Advantage	Disadvantages
Management of the methodology allows us to take into account large numbers of variables.	Cost and implementation time The investment in Artificial Intelligence is very high as they are complex machines with a high cost in maintenance and repair.
Models provide a quick competitive advantage of calibration and re-estimation.	Increase in unemployment. The replacement of humans by machines is leading many people to unemployment on a large scale.
Machine learning favors innovation and the search for new solutions thanks to the interpretation of data.	There is no creativity. Machines do not think, they work within parameters, so the creative capacity remains absent.
Optimized logistics processes will also help us to improve the organization's logistics systems and processes. And it is that it will have a solid database for decision making.	As effective as this technology is, it is not a human being, and it lacks feelings. Thus, as we mentioned earlier, it has no limits and ignores the moral barrier. A circumstance which, if not put on the brakes, can be very dangerous.

Compared to Machine Learning like Deep Learning, they mimic the human brain's way of learning. Their main difference is, therefore, the type of algorithms used in each case, although Deep Learning is more similar to human learning because it functions as neurons. Machine Learning tends to use decision trees and Deep Learning neural networks, which are more evolved. In addition, both can learn supervised or unsupervised.

V. Conclusion and Future Work

In conclusion, the present research work collected accurate information from medical institutions on patients who have ever suffered heart attack problems and on patients who have never suffered such disease, imported libraries for data preparation, data cleaning, and the development of the machine learning model, which in the present case was of the logistic regression type, which gives a result of 1 when there is a presence of probability or 0 when there is an absence of probability.

The data analysis method used was machine learning, which mechanized the construction of our logistic regression model. As a result of the implementation of the model, we had a response of 87% accuracy for people likely to suffer a heart attack and 81% for patients who would not suffer the disease.

As a future topic, it is suggested to implement techniques for data preprocessing such as SMOTE (Synthetic Minority Over- Sampling Technique) for data imbalance. It is also suggested to include in future experiments other variables or characteristics that can facilitate the prediction of the proposed model to optimize it.

REFERENCES

[1] S. S. Khurl and G. Singh, "Ranking early signs of coronary heart disease among Indian patients," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 840-844.

[2] S. Manikandan, "Heart attack prediction system," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 817-820, DOI: 10.1109/ICECDS.2017.8389552.

[3] İ. Berkan Aydilek, "Approximate estimation of the nutritions of consumed food by deep learning," 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 160-164, DOI: 10.1109/UBMK.2017.8093588.

[4] O. Dieste, E. Fernández, R. G. Martínez, and N. Juristo, "Comparative analysis of meta-analysis methods: When to use which?" 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011), 2011, pp. 36-45, DOI: 10.1049/ic.2011.0005.

[5] L. Ibarra, A. Soriano, P. Ponce, and A. Molina, "Research Skills Enhancement through a Research-Based Wit-Learning Methodology," 2019 20th International Conference on Research and Education in Mechatronics (REM), 2019, pp. 1-7, DOI: 10.1109/REM.2019.8744093.

[6] S. Pugh, "Research in engineering, research in design research in engineering design. They're not the same thing," IEE Colloquium on Research in Engineering Design, 1989, pp. 1/1-1/3.

[7] Dua, D. y Graff, C. (2019). Repositorio de aprendizaje automático de la UCI [http://archive.ics.uci.edu/ml]. Irvine, CA: Universidad de California, Facultad de Información y Ciencias de la Computación.

[8] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of a heart attack in coal mining regions using data mining techniques," 2010 5th International Conference on Computer Science & Education, 2010, pp. 1344-1349, DOI: 10.1109/ICCSE.2010.5593711.

[9] N. D. Piza Burgos, F. A. Amaiquema Márquez, and G. E. Beltran Baquerizo, "Métodos y técnicas en la investigación cualitativa. algunas precisiones necesarias, "Conrado, vol. 15, no. 70, pp. 455–459, 2019.

[10] J. Rivera, J. Verrelst, J. Delegido, and J. Moreno, "Herramienta informática para el diseño y evaluación de índices espectrales genéricos para la inversión de parámetros biofísicos."

[11] A. Prieto, A. Lloris, and J. C. Torres, Introducción a la Informática. McGraw-Hill, 1989, vol. 20.

[12] A. Fernández de Castro Fabre and A. López Padron, "Validación mediante método Delphi de un sistema de indicadores para prever, diseñar y medir el impacto sobre el desarrollo local de los proyectos de investigación en el sector agropecuario, "Revista Ciencias Técnicas Agropecuarias, vol. 22, no. 3, pp. 54–60, 2013.

[13] J. Castaño Sánchez, "Análisis y predicción de datos de entrada en urgencias relativos a problemas respiratorios en la ciudad de valencia," 2016.

[14] M. E. Ayala Poma and J. A. Huamán Ollero, "Técnicas y herramientas para la predicción de complicaciones cardíacas, utilizando wearables inteligentes: una revisión sistemática de la literatura," 2020.

[15] David W. Aha (aha '@' ics.uci.edu) (714) 856-8779.

[16] Dua, D. and Graff, C. (2019). UCI Aprendizaje automático Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[17] J. M. Noe, "La potencialidad de la regresión logıstica multinivel: Una propuesta de aplicación en el análisis del estado de salud percibido,"Empiria: Revista de metodología de ciencias sociales, no. 36, pp. 177.

Multi-Criteria Prediction Framework for the Prioritization of Council Candidates based on Integrated AHP-Consensus and TOPSIS Methods

Nurul Akhmal Mohd Zulkefli[1], Mukesh Madanan[3]
Tariq Mohsen Hardan[4]
Department of Computer Science
CAAS, Dhofar University
Salalah, Sultanate of Oman

Muhamad Hariz Muhamad Adnan[2]
Computing Department
FSKIK, Sultan Idris Education University
Tanjong Malim, Malaysia

***Abstract*—Predicting the council candidate becomes difficult due to the large number of criteria that must be known and identified. The best candidate should be chosen from among the candidates because he or she will play an important role in the organization or institution. It is critical to find the right and best candidate these days because people see and judge the outcome from the candidate in a short time with the help of social media. Perhaps the organization and institution require the best candidate criteria because they will manage and organize the community around them. This study focuses on how to prioritize council candidates using Analytic Hierarchy Process (AHP) for determine the criteria and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) for prioritize the student council candidate. This proposed framework based on Multi-Criteria Decision Making (MCDM) will be used to recommend and assist students in selecting the best candidate for student council. The three criteria chosen were grade point average (GPA), Age, and Semester. Based on the results of the questionnaire and a review of the literature, these criteria were developed. The three criteria were then used to determine the most important criterion for selecting the student council. The AHP weight is used to determine and prioritize the most important criteria. TOPSIS was used to select the most qualified student council candidate. The findings show that GPA is the most important criteria in selecting the best candidate, and the TOPSIS findings support the AHP findings.**

***Keywords*—Analytic hierarchy process (AHP); technique for order of preference by similarity to ideal solution (TOPSIS); multi-criteria decision making (MCDM); student council**

I. Introduction

Universities or institutions of higher learning (HLI) provide a variety of responsibilities, such as delivering a high-quality academic curriculum and involving students in extracurricular activities such as sports. It is critical to educate youngsters on the necessary leadership traits for nation-building and future leadership [1]. As a result, student representatives must be included in educational institutions' administrative structures [2]. The student representative committee (SRC) is a student-led organization that aims to foster a feeling of community and leadership among students [3].

Elections are democratic procedures in which the general public chooses a candidate for public office. Elections are crucial to every organization's development because they determine who will lead the people. "A person (a) whose name appears on the official ballot for election to the office of the representative in, or delegate or resident commissioner to, the congress," according to the definition. The student representative council (SRC) is an annual council of university students elected by their peers. Anyone who meets one or more of the following qualifications is eligible for SRC candidate status. Candidates can seek nomination or election via a petitioning method. The candidate can be anybody who wishes to run for office as a write-in candidate. Finally, the candidate also from anybody who selects a treasurer and designates a primary depository, as well as anyone who submits qualifying papers and administers a candidate's oath in compliance with relevant law, is exempt. The goals of this system are to design, implement, and test a web-based system for choosing candidates for the SRC election based on criteria provided by the university.

The proposed system focus on predicting the SRC candidates among students. Differences with other existing systems or methods, most of researchers focus on predicting the result of election. However, predicting the result has become common in the election area. The most important of the election is to get the best and most valuable candidate. The election will finish in one day but the effect of the wrong election result will give a long effect on people, as well as organizations or countries. To determine or predict the best SRC candidate, AHP methods are used to get the best criteria for choosing the candidate. The weight of criteria will be used in TOPSIS to rank and predict the best candidate based on the criteria chosen.

II. Literature Review

When it comes to communicating peers' opinions to the university and, more importantly, ensuring that their opinions are heard, the SRC plays a critical role. They are on the front lines of student welfare, and their opinions are representative of those of students and at the same time, they contribute to the educational environment on campus. Students that join in SRC become participants in the institution's internal decision-making process, enabling them to participate in the governance of the university [3]. The SRC must collaborate with the

university's leadership to ensure that the university's mission and vision are accomplished. SRC members are responsible for student concerns and provide recommendations to the proper university departments, in addition, to serve as the front line of students' advocacy. Therefore, it can be stated that SRC plays an important function in HLI and is a significant stakeholder in the project because of this.

A. Criteria on Student Council Candidate

A strong leader is required for a team to efficiently manage its internal and external affairs as well as to organize teams toward respective objectives [4-5]. The team requires an effective leader who can offer support when necessary and foster interaction as well as trust among team members and also for the people who vote for the candidate. As a result, various research presented several criteria for selecting a strong leader. According to research conducted at Sultan Idris Education University [6], candidates become the primary determinant in being elected as the SRC, followed by their manifesto.

The selection criteria are the exact criteria that candidates must meet to be considered for a job or to fulfill a certain function. Examples of criteria include talents, skills, capacities, and knowledge, among others, but they may also include other characteristics [7]. Apart from that, candidates for the SRC are chosen based on their race, personality, looks, leadership, and academic skills. The research of [8] shows that race and personality are used as selection criteria. Their research revealed that personality traits such as educational attainment, philosophical alignment, and racial affinity were all important considerations in the hiring process, with the respondents giving preference to individuals who met these criteria.

According to the researcher, three qualities for candidates - Commitment, Passion, and Well-organized - received a higher weighting than the other leading criteria. This showed that these three characteristics should be prioritized in the selection of SRC candidates. This research adds to our understanding of the SRC's new vote method, especially in terms of candidate screening. It also offers extensive conditions for SRC candidates, ranking those leadership attributes using the AHP technique [9]. The Student Representative Council (SRC) Election System and the Post Based on Selection Sort (Sorting Algorithm) have one thing in common: every school must have a field of applicants for the SRC position to be chosen by their students. The technique boosts voter participation during elections after a substantial shift, which is one of the benefits.

A project titled A Secure E-Voting for The Student Parliament was created by [10]. The project was built on a cryptographic algorithm and included the pre-election, voting, and election processes. The Student Representative Council (SRC) Election System and the Position Based on Selection Sort (Sorting Algorithm) are comparable in that they are both created for a population of students that is primarily made up of students sharing a common mindset as well as actual behavior. Unlikability, anonymity, and verifiability are all advantages. The drawbacks include increased usability, security, and voter distrust.

Finally, [11] students finished a project titled "Digital Democracy & Student Politics: Interpretation from the Assam University Students Council Election," which focused on how social media might be utilized to convince voters. Both the Student Representative Council (SRC) Election System and the Position based on Selection Sort utilize the internet to encourage students to vote in elections. The benefit is that it gives people a one-of-a-kind chance to express themselves without constraint or interference. The disadvantage is that election management is a difficult undertaking on all fronts.

In this research, we focus on the Ministry of Education Oman's [12] criteria, which include credit hours completed (by semester) and a GPA of at least 2.0 out of 4 points for diploma and bachelor's students, as well as the university council's criteria, which include age as important criteria to consider when selecting candidates.

B. AHP and TOPSIS

Over the last several years, multiple criteria decision making (MCDM) approaches have gained in popularity and are now regularly employed in a broad variety of real-world settings [13-16, 30]. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) proposed by is one of the most popular and widely used MCDM methods [17]. The strategy's underlying principle is straightforward. The so-called positive ideal solution (PIS) and the negative ideal solution (NIS) are used to construct benchmarks (NIS). There were two options considered, and both were chosen because they were closest to the PIS and farthest away from the NIS, respectively. When it comes to benefit and cost, the PIS and NIS are opposites. The PIS optimizes benefit while limiting expense, whereas the NIS maximizes both.

University of Kuala Lumpur students [18] employed the AHP method to increase their instructors' assessment scores. As a part of this study, researchers looked at what criteria contribute to a lecturer's overall performance as well as his or her credibility, as well as how these criteria are ranked based on significance. Because the criteria are weighted depending on importance, the proposed technique is more accurate at differentiating the performance of lecturers than the existing exercise, which takes the average of all criteria into account. Selecting the right SRC is critical because it reflects the good governance of universities [10]. In this study, the Analytic Hierarchy Process (AHP) approach is used to gather student society viewpoints on SRC leadership criteria and to determine which criterion is the most often used.

The Analytic Hierarchy Process (AHP) is used to rank the criteria that demonstrate the importance of dedication, enthusiasm, and organization. Student's curriculum vitae (CV), manifesto plan, and application form are considered as extra requirements in screening the SRC application. The procedure is crucial to ensure that the chosen SRC is competent and capable of fulfilling the expectations of their peers since they will represent the institution [19]. Normally, candidate selection in Malaysian public universities is done through an evaluation process conducted by the faculty/university, but this process is inefficient. Several criteria must be considered to avoid biases. As a result, this study employs the Analytical Hierarchy Process (AHP) method to identify and prioritize the

criteria for selecting the best candidate among UPSI undergraduate students [20]. A common scenario is selecting an appropriate bachelor program and university by [21]. Sijil Tinggi Persekolahan Malaysia (STPM, Malaysian High School Certificate) leavers are a significant group of bachelor program prospect students in Malaysia. Prospective students made their decision based on a variety of criteria, including university requirements, personal preferences, and influences from parents, teachers, and peers. Decisions are typically unstructured and biased as a result of personal preferences and influencers.

Since the emergence of dynamic websites, all business operations of a commercial organization are usually connected with the firm's website. Therefore, a complicated and vast website has been created, which may result in sluggish downloads and difficult navigation. Meeting the demands of the end-user is one of the most fundamental criteria of developing a successful website. Because different users have varied expectations for a website, there are several criteria that the user needs to be satisfied with; hence, evaluating a website is a multi-criteria decision-making problem. The integration of Fuzzy TOPSIS and the Fuzzy Analytic Hierarchy (FAHP) technique to reduce uncertainties and ambiguity in decision making, in which the views of multiple decision makers (DMs) were used for ranking the website [23]. Moreover, another example study is done by [22] that used hybrid fuzzy AHP-TOPSIS to create flood risk maps, hazard, and district-based vulnerability for Istanbul. In health care, example such as [24] examines and evaluates the usefulness of health information systems in the delivery of health care. A multi-criteria study of the efficiency of health information systems utilizing the AHP-TOPSIS approach is used to evaluate electronic health care information systems based on three commonly used software.

AHP and TOPSIS are widely used and well-known in the decision making process. As so far, none integrated MCDM for AHP and TOPSIS is used in prioritizing the council candidates. Most existing research are focused on predicting the result of election compare to help people in deciding on choosing the correct candidates. Besides, predicting the election results are not relevant in this new era as people will look at the candidate's ability to lead the organization or community right after the candidate becomes the leader. So, based on the AHP and TOPSIS method, it can help people to choose the correct SRC candidate based on the criteria given while TOPSIS prioritizes the candidates and help people to choose wisely.

III. Research Methodology

This study had two stages (as shown in Figure1), which were carried out using a mixed-method approach that combined both qualitative and quantitative approaches. The first stage involved evaluating and selecting criteria from the questionnaire and conducting a literature review. Based on this step, three criteria were proposed: GPA, Semester, and Age. Following that, a student sample was collected and stored in the database.

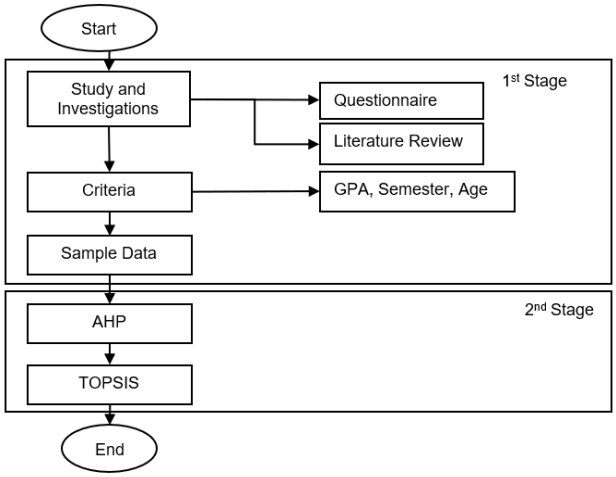


Fig. 1. Framework of Student Council Candidate based on AHP-TOPSIS.

In the second stage, the most important criteria for candidates are determined using AHP, and candidate rankings are based on TOPSIS. The questionnaire and literature review revealed that the criteria for candidates are GPA, Semester (Credit hours), and Age. These criteria were chosen following the discussion in the Literature Review. Meanwhile, sample data were collected by randomly selecting 100 students from a local university to provide input data based on the criteria proposed. After some of the students refused to share the information about their GPA after entering the biography data, the data was cleaned.

In the second stage, eight respondents were chosen to complete the AHP questionnaire to select the best criteria from three options. These respondents were chosen based on their background and experience in either a university setting or an academic setting. We implemented the AHP Balanced-n scale in the AHP method because [25-27] pointed out that integers from 1 to 9 yield local weights that are not evenly distributed.

The AHP indicator was used as shown in Table 1.

TABLE I. Intensity For AHP

Intensity	Definition	Explanation
1	Equal importance	Two elements contribute equally to the objective
3	Moderate importance	Experience and judgment slightly favor one element over another
5	Strong importance	Experience and judgment strongly favor one element over another
7	Very strong importance	One element is favored very strongly over another, its dominance is demonstrated in practice
9	Extreme importance	The evidence favoring one element over another is of the highest possible order affirmation
2, 4, 6, 8 can be used to express intermediate values		

The row geometric mean (RGMM) approach is used to compute the priorities pi in each input sheet Using the *N x N* pairwise comparison matrix $A = a_{ij}$ as calculated in 1

$$r_i = exp\left[\frac{1}{N}\sum_{j=1}^{N} ln\left(a_{ij}\right)\right] = \left(\prod_{i=1}^{N} a_{ij}\right)^{1/N} \quad (1)$$

Consistency index (CI) is supplied with an λ_{max} calculated primary Eigenvalue, either from the RGMM or the principal eigenvalue from the EVM as in 2.

$$CI = \frac{(\lambda_{max}-N)}{N-1} \quad (2)$$

The Alonso/Lamata linear [28] is used to fit the result in CR:

$$CR = \frac{(\lambda_{max}-N)}{2.7699N-4.3513-N} \quad (3)$$

AHP consensus is obtained using Shannon alpha and beta entropy for all inputs. Between zero and one hundred percent, the consensus indicator shows how much agreement there is amongst decision-makers. (Complete agreement among the decision-makers). AHP consensus indicator S^* is calculated using 4.

$$S^* = \left[M - exp(H_{\alpha\ min})/exp\left(H_{\gamma\ max}\right)\right]/\left[1 - exp(H_{\alpha\ min})/exp\left(H_{\gamma\ max}\right)\right] \quad (4)$$

With $M = 1/exp\left(H_{\beta}\right)$. $H_{\alpha,\beta,\gamma}$ is the α, β, γ Shannon entropy for the priorities of all *K* decision makers/participants.

Interpretation of AHP consensus indicator S^* is shown in Figure 2.

Following the AHP process, the sample 100 data will be analyzed using the AHP weight criteria to check if the criteria selected indicate the best candidate. The scores of the trust criteria were ranked in descending order using the TOPSIS technique, whereas the algorithms were rated in the opposite direction. In the TOPSIS technique, the score for each criterion was obtained by calculating the distance between it and the positive and negative ideal solutions. The score with the shortest geometric distance to the positive ideal solution and the largest geometric distance to the negative ideal solution would be the highest utilizing this method. The researcher followed the process of the TOPSIS technique based on [30] in this study.

S^*	Consensus
≤ 50%	Very low
50% - 65%	low
65% - 75%	moderate
75% - 85%	high
≥85%	Very high

Fig. 2. Indicator S^* in AHP.

The first step is to build a normalized decision matrix. To make it easier to compare attributes, certain dimensional attributes were transformed to non-dimensional attributes. Matrix $\left(x_{ij}\right)_{m*n}$ to matrix $R = \left(r_{ij}\right)_{m*n}$ uses the normalization method.

$$r_{ij} = x_{ij}/\sqrt{\sum_{i=1}^{m} x_{ij}^2 r_{ij}} \quad (5)$$

for all i = 1… n and j = 1…n^m

This process yielded a new matrix R, which is shown as follows.

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} \quad (6)$$

Step 2: The weighted normalized decision matrix is constructed.

The weights from decision maker, denoted by $w = w_1, w_2, \dots, w_j, \dots, w_n$ where $j = 1, \dots, n$. Throughout this stage, the normalized decision matrix was used to make decisions. To create the resulting matrix, each column of the normalized decision matrix R was multiplied by the weights in each row of the decision matrix and w_j associated with each column. The weights in the set were all equal to one.

$$\sum_{j=1}^{m} w_j = 1 \quad (7)$$

Step 3: Determination of the ideal and negative ideal solutions

Two artificial alternatives were defined in this phase as A* (the ideal alternative) and A^- (the negative ideal alternative):

$$A^* = \left\{\left(\left(max_i^{v_{ij}|j\epsilon J}\right), \left(min_i^{v_{ij}|j\epsilon J^-}\right)|i = 1, 2, \dots, m\right)\right\}$$

$$= \{v_1^*, v_2^*, \dots, v_j^*, \dots, v_n^*\} \quad (8)$$

Where *J* is a subset of $\{i = 1,2, \dots, m\}$, that has the benefit attributes (i.e., that provides rising utility as the value of I increases), and J^- is the complement set of *J*. Similarly, the cost-type attribute, as expressed by J^c might have the opposite value added as well.

Step 4: Based on Euclidean distance, a separation measurement is calculated as follows. The separation measurement was carried out in this stage by computing the distance between each alternative in V and the ideal vector A* using Euclidean distance, which is given by the equation below:

$$S_{i^*} = \sqrt{\sum_{j=1}^{n}\left(v_{ij} - v_j^*\right)^2}, i = \{1, 2, \dots m\} \quad (9)$$

Similarly, the separation measurement for each alternative in V from the negative ideal A^- is provided by the following equation:

$$S_{i^-} = \sqrt{\sum_{j=1}^{n}\left(v_{ij} - v_j^-\right)^2}, i = \{1, 2, \dots m\} \quad (10)$$

At the end of step 4, two values, namely, S_i^* and S_i^-, for each alternative were counted. The distance between each alternative and the ideal and the negative ideal solutions is represented by these two values.

Step 5: Closeness to the ideal solution calculation

In this step, the closeness of A_i to the ideal solution A^* is defined as

$$C_{i^*} = \frac{S_{i^-}}{(S_{i^-} + S_{i^*})},\ 0 < C_i < 1, i = \{1,2, \dots, m\} \quad (11)$$

Obviously, $C_{i^*} = 1$, if and only if $A_i = A^*$. Similarly, $C_{i^*} = 0$, if and only if $A_i = A^-$.

Step 6: Ranking of the alternatives based on the closeness to the ideal solution. The set of alternatives A_i was now ranked according to the descending order of C_{i^*}, with the highest value indicating the best performance.

A. Result and Analysis

This section shows the outcomes of the individuals who were chosen based on their GPA, Semester, and Age. There are 100 applicants chosen, with 28% between the ages of 18 and 22, 32% between the ages of 28 and 32, and 40% between the ages of 23 and 27. As stated in Table II, 95 candidates are left for the next stage of the procedure after data cleansing.

B. Weighted by AHP

This section presents the data and decision-making results. Three criteria chosen are added to the AHP questionnaire and given to the eight experts. Table III shows the summarization of the respondents' responses.

The Weights acquired from the consensus indicator for group decisions are shown in Table IV. [29] is the one who introduces it. As a result, this group of 8 responders had a consensus S* of 70.6 percent. According to the AHP consensus index, there is a modest level of agreement among participants. The ranking result based on the criteria selected by all respondents (R1 – R8) is shown in Figure 3. Most respondents stated that the most significant criterion for Student Council candidates is a high GPA. The semester is the second most important criterion, and age is the third most critical criterion.

C. Criteria Evaluation by TOPSIS

TOPSIS is used to evaluate the alternatives based on the DM findings presented in Figure 4 and 5, which highlight the sensitivity of the evaluation criteria from the standpoint of each expert.

TOPSIS compares each option to the positive ideal (highest score) and negative ideal (lowest score) by determining the alternative's highest and lowest scoring results (lowest score). S- indicates how near an alternative is to the lowest score. S+, on the other hand, denotes how near an alternative is to the top score. The outcomes of the ranking contexts S- and S+ are shown in Table V.

TABLE II. SAMPLE DATA FOR CANDIDATES

Candidate No.	Candidate GPA	Candidate Semester (Finished hours)	Date of Birth (based on year)	Candidate No.	Candidate GPA	Candidate Semester (Finished hours)	Date of Birth (based on year)
A1	80	20	1980	**A49**	86	56	1986
A2	71	110	1984	**A50**	76	78	2001
A3	83	25	1989	**A51**	78	98	1980
A4	75	35	1988	**A52**	82	70	1999
A5	74	30	1992	**A53**	73	34	1985
A6	80	55	1990	**A54**	65	25	1983
A7	73	34	1989	**A55**	78	92	1996
A8	80	30	1995	**A56**	73	56	1984
A9	89	45	1985	**A57**	77	34	1986
A10	79	78	1992	**A58**	80	77	1995
A11	89	81	1992	**A59**	80	15	2000
A12	70	89	1985	**A60**	81	15	2001
A13	87	60	1990	**A61**	70	65	2001
A14	78	92	1997	**A62**	91	45	1993
A15	79	55	1987	**A63**	72	105	1993
A16	82	111	1996	**A64**	78	60	1998
A17	91	115	1996	**A65**	80	34	2000
A18	88	72	1992	**A66**	87	34	2000
A19	76	102	1993	**A67**	70	89	1999
A20	80	88	1996	**A68**	73	89	2000
A21	84	101	1998	**A69**	82	111	1994
A22	72	67	1993	**A70**	90	90	1993
A23	81	60	1999	**A71**	79	15	2001
A24	79	43	1995	**A72**	68	27	2000
A25	90	88	1995	**A73**	62	83	1992
A26	92	76	1982	**A74**	87	117	1993
A27	92	114	1995	**A75**	75	98	1998
28	74	87	1989	**A76**	80	98	1999
A29	70	28	1988	**A77**	89	60	1993
A30	74	24	1992	**A78**	81	83	1992
A31	82	102	1996	**A79**	78	102	1997
A32	76	22	1996	**A80**	79	80	1997
A33	70	45	1994	**A81**	82	76	1987
A34	72	62	1985	**A82**	73	87	1990
A35	83	76	1991	**A83**	80	93	1992
A36	88	77	2001	**A84**	87	36	2000
A37	72	43	1993	**A85**	73	65	1997
A38	92	102	1995	**A86**	78	67	1992
A39	74	56	1993	**A87**	88	87	1997
A40	93	63	1988	**A88**	86	76	1995
A41	73	110	1993	**A89**	78	102	1992
A42	76	28	1975	**A90**	90	43	1998
A43	89	111	1991	**A91**	83	15	2000
A44	66	66	1993	**A92**	76	78	1994
A45	87	78	2000	**A93**	89	15	2000
A46	88	63	2000	**A94**	67	65	1997
A47	87	54	2000	**A95**	80	50	1994
A48	88	65	2000				

TABLE III. SUMMARY OF RESPONDENTS EXPERT FOR AHP

Experts	Criteria	C_1	C_2	C_3	Weight / Priority	Rank	CR (10%)
1st expert	C_1	1.00	2.00	0.50	28.6	2	0.0
	C_2	0.50	1.00	0.25	14.3	3	
	C_3	2.00	4.00	1.00	57.1	1	
2nd expert	C_1	1.00	0.14	1.00	13.2	3	8.4
	C_2	7.00	1.00	3.00	69.4	1	
	C_3	1.00	0.33	1.00	17.4	2	
3rd expert	C_1	1.00	0.25	5.00	14.3	3	0.0
	C_2	4.00	1.00	2.00	57.1	1	
	C_3	2.00	0.50	1.00	28.6	2	
4th expert	C_1	1.00	4.00	0.25	21.7	2	3.9
	C_2	0.25	1.00	0.11	6.6	3	
	C_3	4.00	9.00	1.00	71.7	1	
5th expert	C_1	1.00	0.50	0.11	7.9	3	1.0
	C_2	2.00	1.00	0.17	14.3	2	
	C_3	9.00	6.00	1.00	77.9	1	
6th expert	C_1	1.00	0.50	0.17	11.7	3	1.9
	C_2	2.00	1.00	0.50	26.8	2	
	C_3	6.00	2.00	1.00	61.4	1	
7th expert	C_1	1.00	0.33	0.11	7.7	3	0.0
	C_2	3.00	1.00	0.33	23.1	2	
	C_3	9.00	3.00	1.00	69.2	1	
8th expert	C_1	1.00	0.50	0.25	14.3	3	0.0
	C_2	2.00	1.00	0.50	28.6	2	
	C_3	4.00	2.00	1.00	57.1	1	

TABLE IV. MATRIX OF NORMALIZATION WEIGHTS

	C_1	C_2	C_3	Normalized Principal Eigenvector	Weights (%)	+/- (%)	Lambda	CR (%)	Consensus
C_1	1	3/4	1/2	22.87%	22.9	1.7	3.006	0.6	70.6%
C_2	1 1/3	1	3/4	32.09%	32.1	2.4			
C_3	2 1/8	1 1/3	1	45.03%	45.0	3.4			

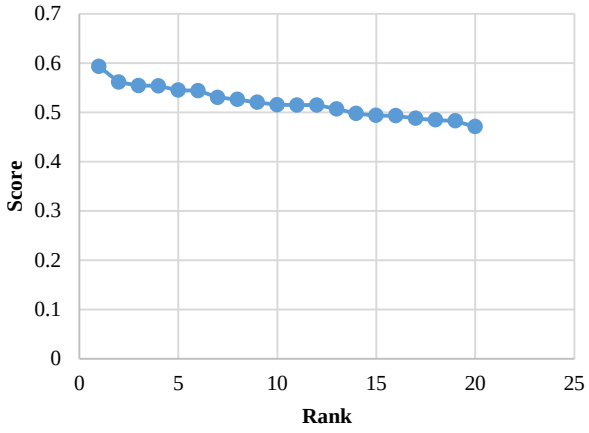


Fig. 3. Score Result for Top 20 Candidates.

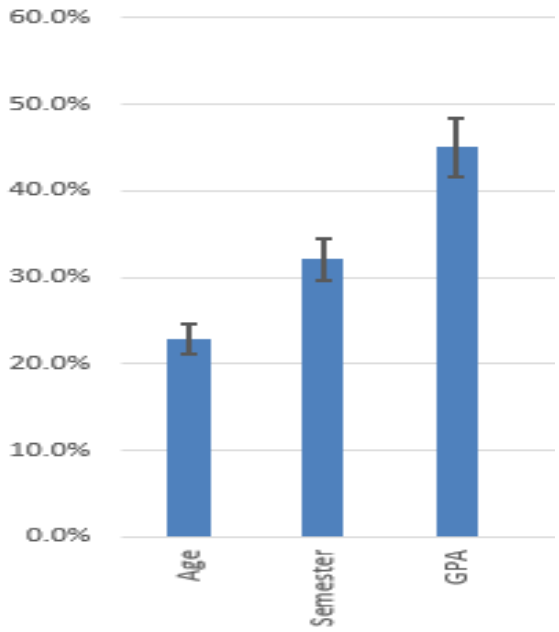


Fig. 4. Percentage Rank based on Criteria.

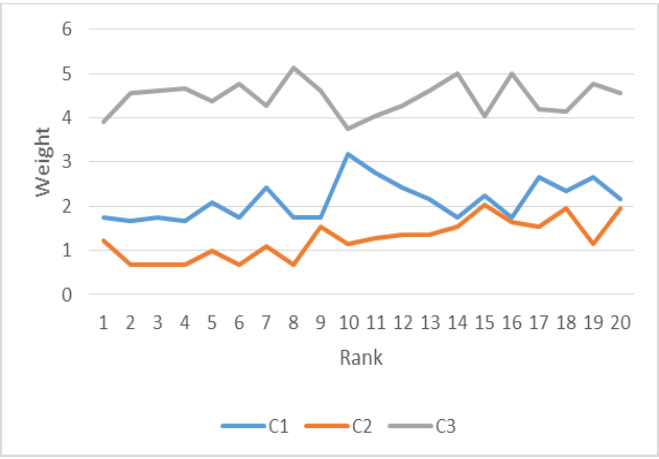


Fig. 5. Criteria Rank based on Weightage.

TABLE V. CANDIDATE'S RANKING RESULT BASED ON THE EXPERT'S WEIGHTAGE

Rank	Candidate	Weight					Score
		22.9	32.1	45	S-	S+	
1	A72	1.751555205	1.225659674	3.920069083	2.296497756	3.342555846	0.59275121
2	A71	1.668147814	0.680922041	4.554197905	3.050229475	3.895731594	0.56086286
3	A59	1.751555205	0.680922041	4.611845979	3.105950806	3.861084259	0.55419315
4	A60	1.668147814	0.680922041	4.669494054	3.159540153	3.861084259	0.55362266
5	A32	2.085184768	0.998685660	4.381253680	2.825427903	3.379200067	0.54462735
6	A91	1.751555205	0.680922041	4.784790204	3.270325871	3.899484783	0.54387556
7	A30	2.418814331	1.089475266	4.265957531	2.765316883	3.121250424	0.53023691
8	A93	1.751555205	0.680922041	5.130678652	3.6014846	3.997681783	0.52606468
9	A65	1.751555205	1.543423293	4.611845979	2.947519616	3.195009582	0.52014608
10	A54	3.169480847	1.134870069	3.747124858	2.619261648	2.784486773	0.51528011
11	A29	2.752443894	1.271054477	4.035365232	2.633837392	2.792470599	0.51461048
12	A5	2.418814331	1.361844082	4.265957531	2.721385221	2.884415107	0.51454178
13	A8	2.168592159	1.361844082	4.611845979	3.001603806	3.083604404	0.50673699
14	A66	1.751555205	1.543423293	5.015382503	3.350595926	3.316096368	0.49741543
15	A33	2.251999549	2.042766123	4.035365232	2.466767121	2.401915148	0.49333884
16	A84	1.751555205	1.634212899	5.015382503	3.348445674	3.253969767	0.49284353
17	A7	2.669036503	1.543423293	4.208309456	2.733085352	2.60017243	0.48753237
18	A37	2.335406694	1.951976518	4.150661381	2.586245751	2.429996051	0.48442621
19	A3	2.669036503	1.134870069	4.784790204	3.316567971	3.092318162	0.48250475
20	A24	2.168592159	1.951976518	4.554197905	2.942836794	2.61721155	0.47071409

IV. CONCLUSION

AHP is effectively employed in computing the objective weight based on the experiment results. The weights are derived from the expert preferences gathered using linguistic criteria. Another issue is that TOPSIS has successfully rated all the candidates based on the user's preferences and the estimated objective weight (see Table IV and Table V). The applicant with the highest rank (A72) obtained the highest score, followed by A71 and A69. The AHP application can help improve the student council selection process by determining the prevailing leadership criteria. When compared to other parameters, the results revealed that GPA, Semester, and Age are the most important.

Because of the successful examination of the AHP and TOPSIS methods, it can be concluded that these approaches can increase the quality of candidates for student council. The study's limitation is the modest size of the student community and expert panel. This research is expected to have a significant impact on future leadership development among organizations and institutions.

REFERENCES

[1] Hamid, J. A., & Krauss, S. E. (2013). Does university campus experience develop motivation to lead or readiness to lead among undergraduate students? A Malaysian perspective. *Journal of Student Affairs Research and Practice*, ***50***(2), 208-225.

[2] Ahiatrogah, P. D., & Koomson, A. K. (2013). Impact of perceived student leadership role on the academic performance of distant education students in Ghana. *The Online Journal of Distance Education and e-Learning*, ***1***(3), 26-34.

[3] Luescher-Mamashela, T. M. (2013). Student representation in university decision making: good reasons, a new lens?. *Studies in Higher Education*, ***38***(10), 1442-1456.

[4] Rou, C. J., Musa, D., & Kamis, N. C. (2017). Students' Awareness towards the Student Representative Council: A Survey Conducted at Northern Region Polytechnics of Malaysia. ***1***(2), 14–22.

[5] Abed Aljasim Muhisn, Z., Omar, M., Ahmad, M., & Adnan Muhisn, S. (2015). Team leader selection by using an Analytic Hierarchy Process (AHP) technique. *Journal of Software*, ***10***(10), 1216-1227.

[6] Boyman, S. N. (2017). Students and Campus Elections: Case study at Sultan Idris Education University, Malaysia. *International Journal of Humanities and Social Sciences*, ***9***(6), 32-45.

[7] O'Meara, B., & Petzall, S. (2009). Selection criteria, skill sets and competencies: What is their role in the appointment of vice-chancellors in Australian universities?. *International Journal of Educational Management*., ***23***(3), 252-265

[8] Mohd Fuad, M.J., Junaidi, A.B., Abdul Halim S., Noor Aziah, M.A., (2011). Persepsi politik belia di kawasan Dewan Undangan Negeri (DUN) Bagan Pinang, Negeri Sembilan. [The political perception of the youths in the state assembly area of Bagan Pinang, Negeri Sembilan]. *Malaysian Journal of Society and Space (Special Issue: Social and Spatial Challenges of Malaysian Development)*, *7*, 105 – 115.

[9] Saaludin, N., Ismail, M. H., Abidin, I. S. Z., & Mat, B. C., (2021). Application Of The Analytic Hierarchy (Ahp) Process For Evaluating Student Representative Committee (Src) Leadership Criteria, *Asia Proceedings of Social Sciences*, ***5***(2), 208-213.

[10] Pilipovic, D. M., & Babic, D., (2016). A secure e-voting for the student parliament, Facta Universitatis, *Series: Electronics and Energetics* ***29***(2), 205-218.

[11] Mishra, R., (2016). Digital democracy and student politics: Interpretation from Assam university student's council elections, The Researcher: International Journal of Management, Humanities and Social Sciences, ***1***(01), 45-56.

[12] MOHE, The Organizational Manual of Student Advisory Councils in Higher Education Institutions, Sultanate of Oman (n.d.) Retrieved August 1, 2021 from https://www.nu.edu.om/contentfiles/SAS-ElectionManual.pdf.

[13] Behzadian, M., Otaghsara, S. K., Yazdani, M., & Ignatius, J. (2012). A state-of the-art survey of TOPSIS applications. *Expert Systems with applications*, ***39***(17), 13051-13069.

[14] Abdullah, L., & Adawiyah, C. R. (2014). Simple additive weighting methods of multi criteria decision making and applications: A decade review. *International Journal of Information Processing and Management*, ***5***(1), 39–49.

[15] Kacprzak, D. (2020). An extended TOPSIS method based on ordered fuzzy numbers for group decision making. *Artificial Intelligence Review*, ***53***(3).

[16] binti Mohd Zulkefli N.A., bin Baharudin B., bin Md Said A. (2018). Trust Blog Ranking Using Multi-Criteria Decision Analysis AHP and TOPSIS. In: Kim K., Kim H., Baek N. (eds) IT Convergence and Security 2017. Singapore.*Lecture Notes in Electrical Engineering*, *450*.

[17] Hwang CL, Yoon K, (1981). Multiple attribute decision making: methods and applications. Springer, Berlin.

[18] Harun, S., Mat, B. C., Ismail, M. H., & Saaludin, N. (2019). Improving Lecturers' Evaluation Score by Using Analytic Hierarchy Process (AHP): A case at Universiti Kuala Lumpur. *Journal Electrical Engineering and Computer Science*, ***15***(1), 391–398.

[19] Saaludina, N., Ismailb, M. H., Zainal, I. S., & Abidinc, B. C. M. (2020). Analytic Hierarchy Process: The Improvement of the Student Representative Committee Selection Method at Universiti Kuala Lumpur. *International. Journal of Innovation, Creativity and Change,* ***11***(12), 719-741.

[20] Mohamed, A. (2021). Criteria For Selection Of The Best Student At Upsi Based On Analytical Hierarchy Process. *Journal of Quality Measurement and Analysis JQMA*, ***17***(1), 93-98.

[21] Yasin, S. N. S., & Adnan, W. N. W. M. (2015). Bachelor Program and University Selection for STPM Leavers using TOPSIS. *Jurnal Teknologi*, ***74***(1).

[22] Ekmekcioğlu, Ö., Koc, K., & Özger, M., (2021). Stakeholder perceptions in flood risk assessment: A hybrid fuzzy AHP-TOPSIS approach for Istanbul, Turkey, *International Journal of Disaster Risk Reduction*. *60*, 102327.

[23] Nagpal, R., Mehrotra, D., Bhatia, P. K., & Sharma, A., (2015) Rank university websites using fuzzy AHP and fuzzy TOPSIS approach on usability, *International journal of information engineering and electronic business.* ***7***(1), 29.

[24] Rađenović, Ž., & Veselinović, I., (2017). Integrated AHP-TOPSIS method for the assessment of health management information systems efficiency, *Economic Themes.* ***55***(1), 121-142.

[25] Salo, A.,Hämäläinen, R., (1997). On the measurement of preferences in the analytic hierarchy process, *Journal of multi-critria decision analysis.* ***6***, 309 – 319.

[26] Goepel, K. D. (2019). Comparison of judgment scales of the analytical hierarchy process—A new approach. *International Journal of Information Technology & Decision Making*, ***18***(02), 445-463.

[27] Goepel, K. (2018). Judgment scales of the analytical hierarchy process: the balanced scale. In *International symposium of the analytic hierarchy process*. Hong Kong.

[28] Alonso, Lamata, (2006). Consistency in the analytic hierarchy process: a new approach, *International Journal of Uncertainty, Fuzziness and Knowledge based system.* ***14***(4), 445-459.

[29] Goepel, K. D. (2013, June). Implementing the analytic hierarchy process as a standard method for multi-criteria decision making in corporate enterprises–a new AHP excel template with multiple inputs. In *Proceedings of the international symposium on the analytic hierarchy process* (Vol. 2, No. 10, pp. 1-10). Creative Decisions Foundation Kuala Lumpur.

[30] Zulkefli, N. A. M., & Baharudin, B. (2015). Travel recommendation system based on trust using hybrid neuro-fuzzy: a study of potential trust in blog and Facebook. *International Journal of Business Information Systems*, *20*(3), 289-309.

A Novel Animated CAPTCHA Technique based on Persistence of Vision

Shafiya Afzal Sheikh, M. Tariq Banday
Department of Electronics and Inst. Technology
University of Kashmir
Srinagar, India

***Abstract*—Image-based CAPTCHA challenges have been successfully used to distinguish between humans and bots for a long time. However, image-based CAPTCHA techniques are constantly broken by hackers, forcing web developers to implement more robust security features and new approaches in CAPTCHA images. Modern-day bots can use many techniques and technologies to break CAPTCHA images automatically. These techniques include OCR, Segmentation, erosion, threshold, flood fill, etc. This led to innovative CAPTCHA systems, including those based on drag and drop, image recognition, fingerprint, mathematical problems, etc. Animated image CAPTCHAs have also been designed to show moving characters and objects and require users to recognize the characters or objects in the animation. Unfortunately, these CAPTCHA systems have also been broken successfully. This research proposes a novel animated CAPTCHA technique based on the persistence of vision, which shows text characters in multiple layers in an animated image. The proposed CAPTCHA technique has been implemented in PHP using GD library functions and tested using various popular CAPTCHA breaking tools. Further, the proposed CAPTCHA challenge has also been tested against the frame separation based breaking technique. The security analysis and usability study have demonstrated user-friendliness, vast accessibility, and robustness.**

Keywords*—*CAPTCHA; OCR; animation; segmentation; botnet; HIP; CAPTCHA usability

I. Introduction

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is some kind of test or challenge which can be solved by a human user very quickly but cannot be solved by modern computer software [1]. These tests help distinguish between humans and computer programs. Unfortunately, hackers widely use automatic computer programs or bots to misuse Internet-based services causing harm to the services and service providers. Therefore, it is incumbent that these services be prevented from automated access and misuse by bots, and it should be done without affecting human users. CAPTCHA challenges help the Internet-based services distinguish between humans and bots, and based on the CAPTCHA test result, and they deny access to the bots. Nowadays, CAPTCHA tests are extensively used on the Internet and have effectively kept away automated bots and prevented misuse of online services only for human users [2].

CAPTCHAs are used for securing web-based services in many ways. They prevent bots from creating email accounts that can misuse online email services and send SPAM emails. CAPTCHAs are also used to prevent search engines and crawlers from accessing web pages and accessing or copying any type of content. Spammers use web crawlers to automatically fetch website content and harvest email addresses from website content. CAPTCHAs can help hide email addresses by default and reveal them only to human users. CAPTCHAs help prevent hacking attacks by defending against brute force and dictionary attacks. These attacks work by trying many login attempts at a fast rate. CAPTCHAs help determine if the client is a human, and only then the service allows further attempts to log in or access the prevented resource. Online services are intended to enable human users to keep trying logging in, even after a bunch of failed attempts, rather than disallow further login attempts, as in the case of bots. Asking a human user to solve a CAPTCHA instead of blocking them offers users a better user experience than being blocked from further login attempts. CAPTCHAs prevent bots from accessing and spamming discussion forums, comment sections of websites, online polling systems and social media applications. Gaming bots can be highly competitive against human users in playing computer games and thus need to be kept away from online gaming platforms. CAPTCHAs help e-commerce websites reject bots that obtain product information and pricing for price comparison. CAPTCHAs play an essential role in keeping these bots away from online services.

Hackers constantly keep trying to bypass CAPTCHAs by implementing CAPTCHA breaking techniques in their bots, making it necessary to design more secure CAPTCHA challenges. There are several types of CAPTCHA challenges currently on the Internet-based services viz. Text-based CAPTCHA, Image-based CAPTCHAs, Audio-based CAPTCHA, Video-based CAPTCHAs, Puzzle-based CAPTCHAs, Mouse based CAPTCHAs and Invisible CAPTCHA.

The most common type of CAPTCHA is Text-based, in which a set of characters is displayed on an image, and the user is required to recognize the text characters and type them in a text box. If the user input matches the characters displayed on the image, the CAPTCHA is accepted as passed. Some examples of Text-based CAPTCHAs are EZ gimpy, Gimpy, Register, Ticketmaster, Yahoo and its multiple Versions, Mailblocks, Google, MSN, Holiday inn priority CAPTCHA, Phpcaptcha.org, FreeCap, Megaupload, BotDetect, Cryptograph, LinkedIn, Authorize, Baidu, Blizzard, CAPTCHA.net, CNN, Digg, Megaupload, Slashdot,

Wikipedia, Hollowstyle, Tencent, Sina, CmPay, MicrosoftCAPTCHA, Baffle Text, TeaBag CAPTCHA 1.2, 3D CAPTCHA, Handwritten CAPTCHA, Synthetic handwritten CAPTCHA, MSN CAPTCHA, 3D CAPTCHA, STE3D-CAP, Sigma-Lognormal CAPTCHA, DevaCAPTCHA, Google CAPTCHA etc.

In Image-based CAPTCHAs, one or more images are displayed to the user, and they are asked to recognize objects on the images. This CAPTCHA type is very effective because computer programs are not smart enough to process and recognize non-text objects from an image. Examples of image-based CAPTCHA are BONGO, Anomalies Image CAPTCHA, Assira CAPTCHA, PIX CAPTCHA, Implicit CAPTCHA, Google Image CAPTCHA, Drawing CAPTCHA, Facebook CAPTCHA, Image Block Exchange, Face Recognition, Multilingual, KitttenAuth, MosaHIP, Image Flip CAPTCHA [3] etc.

Audio-based CAPTCHA is another CAPTCHA type, but they are not as common as the other CAPTCHAs. They allow users to play a sound and recognize words spoken in the audio. The audio usually has background noise to prevent it against voice recognition based breaking attacks. This CAPTCHA type helps blind or visually impaired computer users pass CAPTCHA tests. Examples of audio-based CAPTCHA are CAPTCHA for blind users, HIPUU, Google reCAPTCHA, Digg etc.

Video-based CAPTCHAs are yet another type of CAPTCHAs that are rarely used. They require a user to watch a short clip and then answer a question based on the information provided in the video. E.g., recognizing a human gesture, moving objects or text from the video. Video-based CAPTCHAs require more internet bandwidth and use attentiveness and time. Some examples of video CAPTCHA are 3D animation CAPTCHA, AniCAP, Motion CAPTCHA, New video CAPTCHA, NuCAPTCHA, HelloCAPTCHA, DotCHA, etc.

Another common type of CAPTCHA scheme is Puzzle-based, in which the user to require to solve a small, easy puzzle that a computer program cannot solve. It depends solely on human intelligence because computer programs are nowhere near good at solving random puzzles. A few examples of puzzle CAPTCHA are 3D animation CAPTCHA, AniCAP, Motion CAPTCHA, New video CAPTCHA, NuCAPTCHA, HelloCAPTCHA, DotCHA, etc.

Mouse based CAPTCHAs are effective and very easy to use. They require users to click a button or a checkbox to declare that they're not bots. The CAPTCHA system records the users' mouse movement patterns and analyses those patterns to determine whether or not the user is a bot. The best example of this type of CAPTCHA is Google reCAPTCHA v2. Other examples are Mouse CAPTCHA, unCAPTCHA, Drag and Touch CAPTCHA [4] etc.

Invisible CAPTCHAs are gaining popularity lately. They do not require users to do anything, making them the most user-friendly way to distinguish between humans and computer programs. They work by analyzing users' previous actions like recent website activity, session information and other parameters like browser or client information, IP address reputation etc. The best example of invisible CAPTCHA is Google reCAPTCHA v3 [5].

CAPTCHAs are not immune to attacks. Hackers constantly keep trying to break or bypass CAPTCHA systems to perform their activities efficiently. They make use of advanced techniques to find ways to break the CAPTCHAs. CAPTCHAs in all of the types mentioned earlier have been successfully broken.

Simple Text-based CAPTCHA challenges are very easy for a bot to break with the advent of image segmentation and OCR technology. To prevent text-based CAPTCHAs from being broken using OCR and related technologies, the text characters on the image are distorted and deformed in different ways to make automatic text recognition difficult while still keeping them recognizable by human users [6], [7].

Image-based CAPTCHA challenges have also been broken successfully using feature extraction (colour and texture), SVM classification techniques, face detection using kNN classification techniques, google reverse classification, HSV model, image collection, tag classification techniques. Audio-based CAPTCHAs have been broken using vertical segmentation, DFT recognition, Ada Boot, SVM, CNN techniques.

Video-based/Animation CAPTCHAs have also been broken using frame selection, pixel display timing, vertical segmentation, connected pixels, flood fill, k-means clustering techniques using SIFT and NN classifications.

Mouse-based CAPTCHA has been successfully broken using fake click implementation techniques, image annotation services and tag classifiers. In addition, invisible CAPTCHAs have also been broken using Reinforcement Learning techniques with a high success rate [5], [8], [9].

A. Contribution

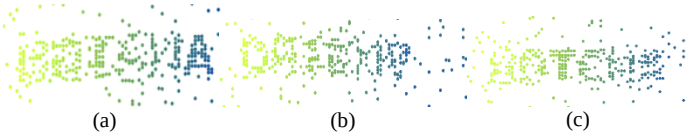
In this research, a CAPTCHA has been designed, which works on the concept of persistence of vision or retinal persistence. On the retina of an eye, the visual perception of an object does not end immediately and remains for a fraction of a second even after the light coming from it stops entering the eye. This research proposes to use this phenomenon of the human eye to display a set of images to the users at a fast frame rate, each frame showing partially visible alphanumeric characters, giving the user an illusion of completely visible numbers or characters. Furthermore, the proposed CAPTCHA displays two sets of partially visible characters, one after another, for a short duration. Individually, none of the images shows any of the characters thoroughly, making it difficult to extract the character information from individual frames of the animated CAPTCHA. Therefore, the CAPTCHA is very secure against frame separation, segmentation and OCR based CAPTCHA breaking techniques.

II. Literature Review

AniCap is an animated 3D text-based CAPTCHA challenge based on the concept of motion parallax. Each 3D character in AniCap has a random 3D transformation that rotations in all three dimensions. Unlike other approaches that add random

clutter to the CAPTCHA challenge to deter automated attacks, it uses overlapping text-on-text with no distinct colours or borders around the character. The two layers of text move around at different paces that give the user a feeling that the two layers of text are to varying distances from the users' perspective. This allows the user to recognize the text in front and back layers separately. AniCap is difficult to break using common CAPTCHA breaking techniques because the foreground and background colours are not distinct, and therefore edge detection is not possible easily [10].

DotCHA is a new and unique 3D animated CAPTCHA that displays text and uses human interaction to avoid the shortcomings of existing 2D and 3D CAPTCHAs. DotCHA requires users to use a mouse or finger gestures on a mobile device to rotate a random-looking extensive collection of small balls, constantly moving in a 3D circular direction, a 3D text model, to identify the correct letters. The 3D model formed by the collection of moving balls is a twisted form of 3D letters around a centre pivot axis, and it displays different letters at different angles of rotation. This is because the balls line up to form a shape of letters only at specific angles of rotation. Because the characters in DotCHA are made up of many small balls instead of solid colour text characters, it belongs to the scatter-type CAPTCHA category. Therefore, it can't be broken using segmentation techniques. DotCHA is also safe against machine learning-based attacks because the characters are recognized only at a specific degree of rotation [11]. However, our analysis of DotCHA was very user-unfriendly because it is tough to recognize the characters as they are not distinctly distinguishable due to a lack of solid colour and distinct borders or edges. Fig 1. shows a few screenshots of DotCHA.



(a) (b) (c)

Fig. 1. DotCHA Sample Images. (a) in which the Character 'A' is Visible. (b) in which the First Character is Supposed to be 'D' (c) the Letter 'T' is Barely Recognizable.

HelloCAPTCHA is an animated CAPTCHA web service that generates CAPTCHA animations on the HelloCAPTCHA server. It generates 2D animations which display animated text characters. The service generates at least 84 known different variants of animations. HelloCAPTCHA provides programming interfaces for PHP and JAVA web applications and plugins for Joomla, Drupal, WordPress. (hellocaptcha.com). HelloCaptcha Developers claim that their animated CAPTCHA schemes are user-friendly, secure against breaking attacks, especially because of the extra time required to break the CAPTCHAs over multiple frames instead of a single image in traditional static CAPTCHAs [12]. [13], [14] have successfully broken HelloCAPTCHA using various techniques that include Pixel Delay MAP (PDM), Calculating Line, Color Selection, Frame Selection etc. They report a success rate between 16% and 100%.

Motion CAPTCHA is a video-based CAPTCHA scheme in which a short video clip is displayed to the user. The video shows a human performing some action or gesture. The user must watch the clip, recognize the action or gesture performed in the video, and choose the correct description of the action from a list of actions provided to the user alongside the video. This CAPTCHA scheme requires users to be good at understanding the English language and takes time to watch the video. The video may also take time to download in case of less internet bandwidth. [15] highlights the weaknesses of the MotionCAPTCHA.

Gesture-based animated CAPTCHA is a unique type of animated CAPTCHA in which sign language is used to convey numbers utilizing a video clip. For example, in the video clip shown in Fig.2., hands are rendered, which offer a few numbers in sign language. This type of CAPTCHA is easy for users familiar with sign language, but it is initially difficult for those who don't know sign language, which is the majority of the people in the world. Furthermore, this CAPTCHA type can't be broken using conventional CAPTCHA breaking methods and tools. Instead, it will require an AI-based system to recognize the hand gestures and break such CAPTCHA challenges [16].

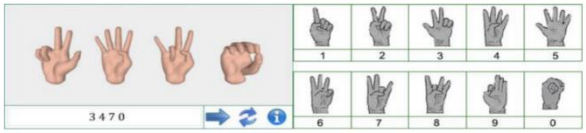


Fig. 2. Shows some Sample Hand Gestures Rendered using this Technique and Instructions for Understanding some Numbers in the Sign Language.

III. Proposed Solution

This research proposes creating an animated CAPTCHA challenge designed to mitigate the security issues that make animated CAPTCHAs insecure against breaking techniques. The proposed animated CAPTCHA is based on image retention by the retina of the eye for a short duration after the image is not visible anymore. Because the human is able to process the data in such a way that it has the ability to create meaning from a sequence of meaningless partially visible pieces of an object, seen in a sequence for a very brief period of time, it is possible to use this special ability to distinguish a human from a machine which doesn't have such an ability. This special ability of the Human eye can be used to create an illusion of fully visible text characters, even though they are shown piece by piece, sequentially at a fast speed.

The CAPTCHA system generates a set of images with random background colours, lines and patterns. A transparent image is then created on which a group of randomly generated alphanumeric characters is printed. The number of characters is chosen randomly for the two sets of frames. The alphanumeric characters are printed in a random colour and font size between 12 and 16. The text is printed starting at the random top and left pixel locations and is also rotated randomly between +30 and -30 degrees before being printed on the image.

Then, sequentially, parts of all the characters are erased from the image to create a frame with partially visible characters. This process is repeated several times to create a set of frames, and in every frame, different parts of the characters are trimmed off. These transparent frames are then pasted on the previously generated background images. Creating the

frames is repeated with a different set of alphanumeric characters to create another set of frames. Hence, two sets of frames are constructed with two different sets of alphanumeric characters on them. The two sets contain different frames every time a new CAPTCHA is generated, so the number of frames for each character set is not fixed. Finally, the two sets of frames are combined to form a gif animation, completing the process of generating the CAPTCHA challenge. The output is an animation that shows a group of animated characters for a few seconds and then replaces the characters with another set of characters for the next few seconds. The flowchart of the proposed CAPTCHA scheme is shown in Fig. 3.

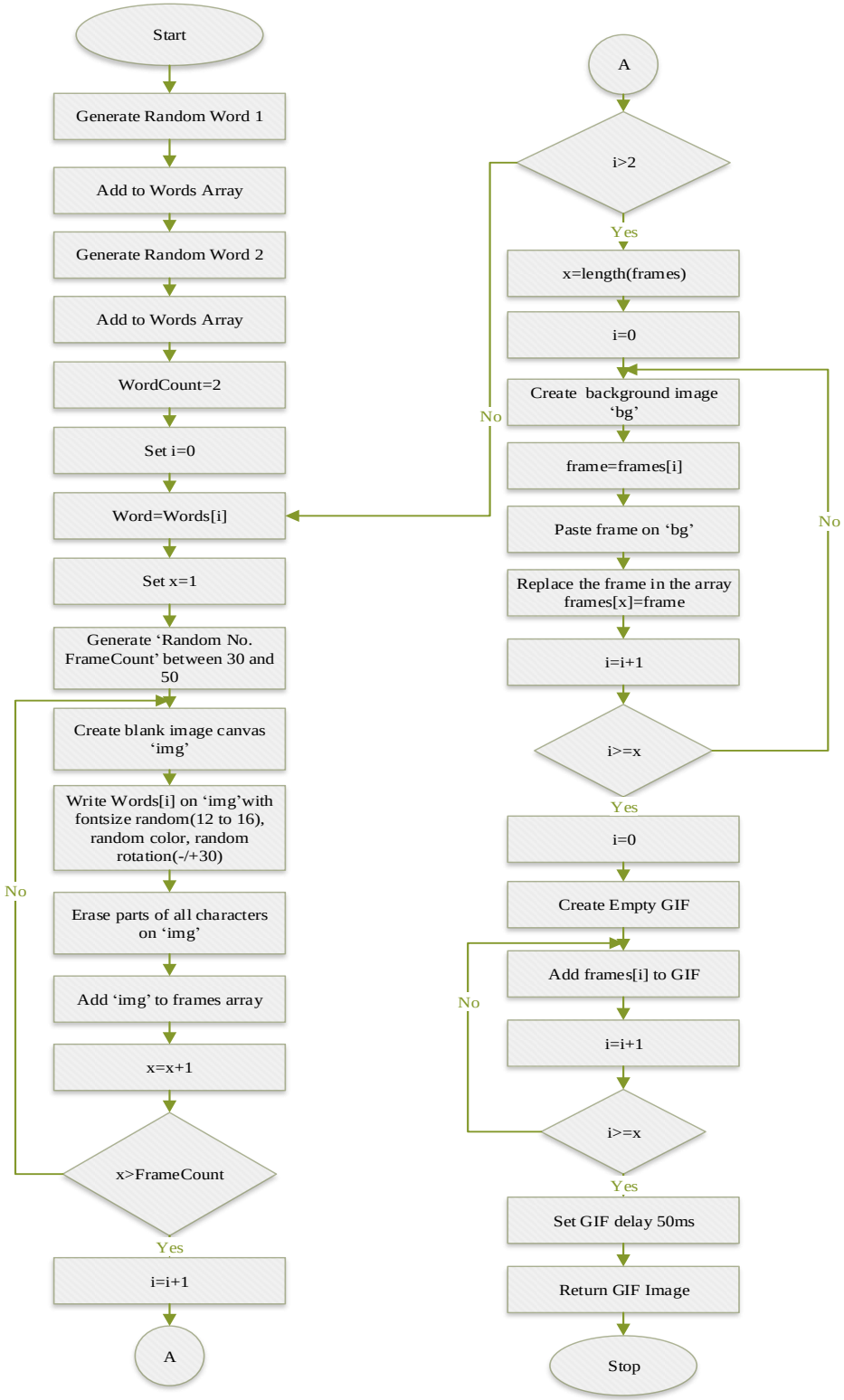


Fig. 3. The Flowchart shows the Implementation of the Proposed CAPTCHA Scheme.

The flowchart starts with generating two words or sets of random characters that will be displayed in the proposed CAPTCHA. The two words are stored in an array. Then, for each word, an integer variable FrameCount is created and its value is set randomly between 30 and 50. Then a blank/transparent image is created and the first word in the array is written on the image using a random font size between 12 and 16 using a randomly created colour and the text is printed on the image at an angle, the degree of which is chosen randomly between -30 and +30 degrees. Then parts of the image are cleared so that the text is not readable. After that, the image is pushed into an array named Frames. This process is repeated FrameCount times, thereby creating a random number of frames, between 30 and 50. This process is repeated for the second word in the array of Words. This results in the creation of two sets of frames for two randomly chosen sets of characters. Both sets contain between 30 and 50 frames. Then, for each frame in the Frames array, a random background image is created with some random noise and colours. One by one, the two sets of frames are pasted on top of the random backgrounds, resulting in a set of frames containing partially visible text characters on random background images. Finally, all the frames are joined together and converted into a GIF image with a frame delay of 50ms between the frames. The resulting GIF image is output and is the desired CAPTCHA scheme proposed in this research.

IV. Implementation

The proposed CAPTCHA has been implemented and tested in this research. The implementation was done in PHP language using GD library functions [17]. The GD library allows creating blank images using the imagecreatetruecolor() function on which the random text characters are printing using imagettftext() function, with random text colour and random font size, from a pre-determined size range at a random angle of rotation. The top and left pixel location for printing the text is also chosen randomly for every set of frames. The colour, font size and rotation angle and location are passed as parameters to the imagettftext() function. Then, using the imagecolortransparent() function, the system is instructed to render a pre-defined RGB colour (Ct) transparently. After that, parts of the text characters pasted on the image are removed and made transparent by adding rectangular shapes of Ct colour over the text. Fig. 4. shows some of the frames with partially visible text:

The frames thus created are pasted on randomly generated background images using the imagecopy() function. The frames are then combined and encoded as a gif image producing the final animated CAPTCHA.

Fig. 4. (a) shows a simple text printed on a blank image and (b) to (f) show a few frames with different parts of the text trimmed off. It can be noticed that the text is not visible or barely recognizable in these frames. Fig. 5. shows a few frames of the text over random background images, which also do not have fully visible text characters. It is impossible to show a screenshot of the CAPTCHA in motion, the way the human eye perceives it because a screenshot can only show a single frame.

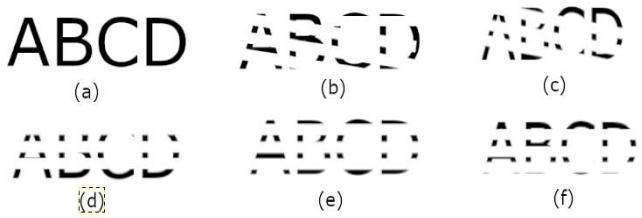


Fig. 4. A Sample Actual Text Frame and Frames with Partially Visible Text.

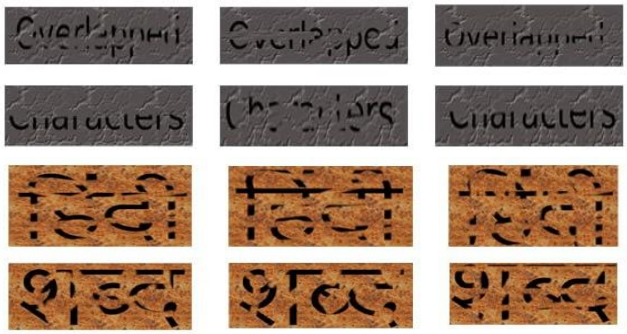


Fig. 5. Sample Frames from the Final CAPTCHA Animation.

The proposed CAPTCHA has been implemented and tested in English and Hindi languages and can be implemented in any other language. Furthermore, the CAPTCHA scheme was integrated and tested with a Multilingual CAPTCHA system [18], [19], [20], wherein the proposed CAPTCHA could create challenges in all the supported languages.

V. Security Features and Analysis

The animated CAPTCHA can be broken down into individual frames, and the individual frames can be subject to various captcha breaking algorithms, including segmentation, etc. The motion CAPTCHA is challenging to break using OCR or AI techniques as the recognition of characters from animation is difficult. Separate frames of the gif are useless for the OCR script. Background and noise randomization increase the problem for bots. The fast frame rate makes the characters seem intact, while in reality, the characters are NOT complete in ANY of the frames separately. The characters are incomplete, and therefore, moderate character distortion will suffice. Further, the presence of two sets of characters overlapped in the same place doubles the challenge of breaking the CAPTCHA because the gif image contains parts of two different characters at the same place, overlapped, in different frames. So, the challenge for any bot trying to break the CAPTCHA is to recognize two overlapped, partially visible characters from the same location in the various frame. If a bot tries to obtain multiple parts of a character from a given location, from multiple frames, it will have to figure out which parts belong to which character set. The following attacking techniques have been considered while designing the proposed CAPTCHA:

A. Segmentation

Segmentation techniques are used to break the text on a CAPTCHA image into various characters. The proposed CAPTCHA uses rotation and change of position of characters,

the background and foreground sets of lines—also, the entire set of characters' changes somewhere at the middle of the animation. The partially trimmed characters make the text on individual frames unrecognizable, even to the human eye. These features make the segmentation of characters extremely difficult, if not impossible.

B. Number of Characters

Common breaking techniques start by counting the number of characters on a CAPTCHA image and dividing the image into that many pieces. This makes it easy to extract a single character from a piece of the image. This research suggests using a different number of randomly chosen characters every time the CAPTCHA is requested. This makes it difficult to predict the number of characters on a CAPTCHA image, and thus breaking the image into pieces doesn't guarantee a single character on every part of the image. Furthermore, it is suggested to use a different number of characters for the two sets of characters proposed in this research. The varying font size on other framesets and CAPTCHA images further complicate the attack process.

C. Position

The starting position of the text on the proposed CAPTCHA varies every time, and over multiple frames, the characters don't stay at a single position but keep moving slightly and slightly, changing the rotation. This way, predicting the position of characters isn't possible in the proposed CAPTCHA, which mitigates a range of attacks.

D. Proper use of Color

Incorrect use of colours can adversely affect the security of any type of CAPTCHA. This research proposes multi-colour background objects and lines so that reducing the number of colours in frames doesn't reveal helpful information about the background and text colours.

E. Number of Frames

The proposed CAPTCHA has a different number of frames every time a CAPTCHA is generated. As the proposed CAPTCHA displays two sets of characters during the animation, it is possible to predict that a different set of characters becomes visible halfway, i.e. after 50% of the frames. Therefore, the proposed CAPTCHA distributes the frames among the two sets of characters unevenly, using a different number of frames per character set. Furthermore, the distribution varies per CAPTCHA generated; therefore, the two-character sets don't have the same number of frames. This makes it difficult for any attack to determine which frame the character set changes.

F. Frame Delay

The frames in the proposed CAPTCHA are delayed unevenly. The frame rate varies randomly between 1 and 10 milliseconds for every frame. This complicates the prediction of the rate of change of various pixel characters for any breaking attack.

G. PDM Attack

The attack based on Pixel Delay Map (PDM) work assumes that the characters required to be recognized by human users will be displayed for a longer duration of time than other moving parts of an animated CAPTCHA. The pixels that don't change colour for a fixed period are mapped to determine the location of the text characters. The proposed CAPTCHA doesn't have any moving background or foreground pixels. The entire background and partially visible sections of the foreground text are static and visible for the whole duration of the frame. Therefore, the PDM technique of breaking animated CAPTCHAs is useless on the proposed CAPTCHA scheme.

The security analysis of the proposed CAPTCHA was done using GSA CAPTCHA Breaker [21] and CAPTCHA Sniper [22], which are very popular and powerful CAPTCHA breaking tools. First, a set of 200 CAPTCHA animations was generated using the implementation of the proposed system, which was used for the security analysis. Then, one by one, the animations were broken down into individual frames or images. Each animation was broken down into 60 frames, resulting in 12000 frames. All the 12000 frames were analyzed individually using CAPTCHA Sniper and GSA CAPTCHA Breaker with a varied set of configurations and combinations of configurations. Because the proposed CAPTCHA scheme doesn't expose text character information in individual frames, these tools did not recognize a single character from 12000 frames.

The only way to break the proposed CAPTCHA is to collect individual text character information from more than a few consecutive frames. First, the selected frames must be processed to clear the background from the frames and leave the visible sections of the text untouched. Then the parts of the text characters must be extracted from individual frames and combined in a single image to form recognizable text characters. The proposed CAPTCHA mitigates this type of attack by slightly changing the location of the text by a few pixels in every frame. However, the change of location is performed randomly in the x and y axes. Therefore, if collected over multiple frames and combined, the parts of the text will result in a heap of small chunks of characters, completely unrecognizable, as shown in Fig. 6. The proposed CAPTCHA was tested against the breaking method discussed above. To simplify the attack, the step for adding the background image to the frames was skipped when generating the CAPTCHA, and the individual frames were exported to transparent backgrounded PNG images. The first step in breaking the CAPTCHA by removing the background objects and noise was not required. The frames thus contained partially visible text characters on a clear transparent background. The frames were then superimposed on each other to combine the partially visible parts of the text, expecting to get the full-text characters. The resulting image was then read using Tesseract OCR [23], [24], which didn't succeed in recognizing any characters.

Fig. 6. Sample Actual Character and Character Information Combined from 15 Frames.

VI. USABILITY

Unlike traditional static image-based CAPTCHA, the text characters in the individual frames in the proposed CAPTCHA don't need any transformation, deformation, or visual effects to make it difficult for bots that use various text extraction and recognition algorithms and OCR technologies to recognize the characters. It is an illusion created by the animation in the human brain that the text becomes visible and recognizable to the humans. At the same time, in reality, the frames are entirely useless for machines or bots. The user, therefore, doesn't have to work hard and try to recognize deformed text, as in the case of most types of CAPTCHAs.

An online usability study was carried out to test the usability and user-friendliness of the CAPTCHA images. Five hundred volunteers were asked in different social media groups to solve the proposed CAPTCHA challenge, with each user asked to solve five challenges. A web page was created, in which the user was asked about their preferred language, the options being English and Hindi. Based on the language chosen, the users were shown a second page with five CAPTCHA challenges in their chosen language, each challenge having a textbox for entering the answer. The responses were verified and the answers stored in the backend database. As shown in the Table 1, five (5) English language CAPTCHAs were attempted by 350 English-knowing users, which is a total of 1750 attempts. Out of the 1750 attempts, 1732 were correctly solved and 18 attempts were unsuccessful. For Hindi, 5 CAPTCHA challenges were solved by 150 users which sums up to a total of 750 attempts. Out of the 750 attempts, 741 were correctly solved and 9 were failed attempts. In combination of both languages, 2473 attempts were successful out of a total of 2500 attempts, giving 98.92% success rate.

TABLE I. USABILITY STUDY OF PROPOSED CAPTCHA

Language	**No. of Users**	**Challenges per User**	**Total Attempts**	**Solved**	**Failed**	**Success %age**
English	350	5	1750	1732	18	98.97
Hindi	150	5	750	741	9	98.8
Total	**500**	**5**	**2500**	**2473**	**27**	**98.92**

VII. CONCLUSION AND FUTURE SCOPE

Animated CAPTCHAs are an effective way to keep bots away and prevent online services from being misused. But they are vulnerable to breaking attacks and are broken easily by analyzing individual animation frames. This work discusses various available animated CAPTCHAs and techniques that have been used to break them. Animated CAPTCHAs, however, have the potential to do that can be done using static images. Therefore, a new animated CAPTCHA scheme has been proposed and implemented that makes use of the animation techniques and the natural behaviour of the retina of an eye to propose a very effective CAPTCHA scheme, which is immune to the traditional animated CAPTCHA breaking attacks and techniques. The CAPTCHA has been designed keeping the security loopholes in mind that weaken the animated CAPTCHAs. The proposed CAPTCHA has been tested against possible attacks programmatically using GSA CAPTCHA Breaker and CAPTCHA Snipper. The proposed CAPTCHA is user-friendly, easy to use, secure and innovative, and easy to implement. The traditional animated CAPTCHAs display the characters in only specific frames, thereby making users wait and spend quite a bit of time trying to recognize the characters. The proposed CAPTCHA text is visible to the user throughout the animation cycle. It does not require any special browser plugins to display the animation because it is created as a gif animation.

The phenomenon of image retention of the retina of Human eye is an extremely unique feature which has been used in this research to distinguish a human from a computer by implementing and successfully testing a novel CAPTCHA challenge based on its special capability. Therefore, we believe that this concept opens up possibilities for future research towards designing novel, highly secure and user friendly CAPTCHA challenges using this phenomenon.

ACKNOWLEDGEMENT

This work has been supported by Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India under its file no. EMR/2016/006987.

REFERENCES

[1] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford. "CAPTCHA: Using Hard AI Problems for Security," InE. Biham, editor, EUROCRYPT, volume 2656 of Lecture Notes in Computer Science, pages 294–311. Springer,2003.

[2] K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski. "Building Segmentation Based Human-Friendly Human Interaction Proofs (HIPs)," In H. S. Baird and D. P. Lopresti, editors, HIP, volume 3517 of Lecture Notes in Computer Science, pages 1–26. Springer, 2005.

[3] M.T. Banday, N.A. Shah, N.A, "Image Flip CAPTCHA," ISeCure, The ISC International Journal of Information Security, Iranian Society of Cryptology, Tehran, Iran, ISSN 2008-2045 and 2008-3076, 1(2), pp. 103-121, 2009.

[4] A.R.Shah, M. T. Banday, S.A.Sheikh. "Design of a Drag and Touch Multilingual Universal CAPTCHA Challenge", Department of Computer Science & Engineering, ABES Engineering College, Ghaziabad on 22-23 Feb 2019, Springer.

[5] M.Kumar, M. K. Jindal and Munish Kumar, "A Systematic Survey on CAPTCHA Recognition: Types, Creation and Breaking Techniques," Archives of Computational Methods in Engineering, Springer, June 2021

[6] O. Starostenko, C. Cruz-Perez, F. Uceda-Ponga, et al., "Breaking text-based CAPTCHAs with variable word and character orientation". Pattern Recognition, 2015, 48(4): 1101-1112.

[7] G. Mori, J. Malik. "Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA," International Conference on Computer Vision and Pattern Recognition Proceedings, Washington, 2003, 134-141.

[8] I.Akrout, A.Feriani, M.Akrout, "Hacking Google reCAPTCHA v3 using Reinforcement Learning," arxiv: cs.LG/1903.01003.

[9] S. Sivakorn, I. Polakis, A.D. Keromytis, "I am Robot: (Deep) Learning to Break Semantic Im-age CAPTCHAs," In 2016 IEEE Eur. Symp. Secure. Priv. (EuroS P), pp. 388–403 (2016). DOI10.1109/EuroSP.2016.37.

[10] W. Chow, W . Susilo W, "AniCAP: an animated 3D CAPTCHA Scheme based on motion parallax," In: Proceedings of 10th international conference on cryptology and network security, pp 255–271, 2011.

[11] S. Kim, S. Choi, " DotCHA: A 3D Text-based Scatter-Type CAPTCHA", Web Engineering; Bakaev, M., Frasincar, F., Ko, I.Y., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 238–252.

[12] http://www.hellocaptcha.com/; accessed on May 2021.

[13] V. Nguyen, Y. Chow, W. Susilo, "Breaking an animated CAPTCHA scheme", In International conference on applied cryptography and network security, 2012, pp 12–29.

[14] S. S.A. Shah, R. A.Shaikh, R. H. Arain, "Reading the Moving Text in Animated Text-BasedCAPTCHAs," In International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 12, 2018.

[15] L.A. Leiva, F. Álvaro, µcaptcha: "Human Interaction Proofs Tailored to Touch-Capable Devices via Math Handwriting," In: International Journal of Human-Computer Interaction,2015, 31(7), pp.457–471.

[16] A. Shumilov, A. Philippovich, "Gesture-based animated CAPTCHA", In Information and Computer Security; Bingley Vol. 24, Iss. 3, pp. 242-254, 2016.

[17] S. Stobart, M. Vassileiou, "GD Library," In PHP and MySQL Manual. Springer Professional Computing. Springer, London, (2004).

[18] M.T. Banday, S.A. Sheikh, "Design of CAPTCHA Script for Indian Regional Websites," In Communications in Computer and Information Science Security in Computing and Communications, Springer Berlin Heidelberg, pp. 98–109, 2013.

[19] M.T.Banday, S.A. Sheikh. "Design of Secure Multilingual CAPTCHA Script", International Journal of Web Portals, IGI Global Vol. 7, No. 4, pp. 1-27, 2015.

[20] M.T.Banday, S.A. Sheikh. "Service Framework for Dynamic Multilingual CAPTCHA Challenges: IN-CAPTCHA" 2014 International Conference on Advances in Electronics, Computers and Communications (ICAECC-2014) 10-11 October 2014, Reva Institute of Technology and Management, Bangalore, India, published by IEEE, ISBN:9781479954971, pp. 1-6.

[21] GSA CAPTCHA Breaker, "https://www.gsa-online.de/product/captcha_breaker/", accessed on October 2021

[22] Captcha Sniper (CS), "https://www.captchasniper.com/", accessed on December 2021.

[23] Tesseract OCR, " https://github.com/tesseract-ocr/" accessed on Nov. 2021

[24] R. Smith, An Overview of the Tesseract OCR Engine, In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), IEEE Conference, 23-25 September 2007, Brazil.

Securing Dynamic Source Routing by Neighborhood Monitoring in Wireless Adhoc Network

Rajani K C[1]
Assistant Professor
Department of Computer Science and Engineering, Sea College of Engineering and Technology, Bangalore, India

Aishwarya P[2]
Professor and HoD
Department of Computer Science and Engineering
Atria Institute of Technology, Bangalore, India

***Abstract*—Wireless Adhoc Network (WANET) significantly contributes to cost-effective network formulation due to decentralized and infrastructure-less schemes. One of the primary forms of WANET in Mobile Adhoc Network (MANET) is still evolving in research and a continued set of research problems associated with security. A review of existing security approaches shows that identifying malicious behavior in MANET is still an open-end problem irrespective of various methods. This paper introduces an improved DSR protocol mechanism of neighborhood monitoring scheme towards analyzing the malicious behavior in the presence of an unknown attacker of dynamic type. The proposed method contributes to deploying auxiliary relay nodes and retaliation nodes to control the communication process and prevent the attacker from joining the network. Using analytical research methodology, the proposed system can offer better communication performance with effective resistance from threats in MANET.**

Keywords*—Mobile adhoc network; wireless adhoc network; security; attack; dynamic source routing

I. Introduction

The current research work is focused on deploying a security feature for identifying the threat followed by preventing the threat in Wireless Adhoc Network (WANET). Used in multiple forms of wireless application, adoption of WANET offers better connection as well as it also invite various levels of threats. Basically, a WANET is a decentralized scheme used in the wireless network with zero dependencies towards any infrastructure [1]. Theoretically, it is further classified into Wireless Sensor Network (WSN), Mobile Adhoc Network (MANET), and Wireless Mesh Network (WMN). However, from a practical viewpoint, the mapping of WANET is more carried out towards the MANET environment characterized by various constraints, e.g., low resource availability, lower processing capability, decentralized, and dynamic topology [2]. Besides, establishing a potential route among mobile nodes in a dynamic environment irrespective of various routing-based research approaches [3]. Owing to this limitation in MANET, there is a likely probability for any mobile node to reject forwarding data packets and thereby act as a malicious node. At present, there are various forms of security threats in MANET, and ongoing research is attempting to solve this problem to a maximum extent using different approaches [4][5]. Out of all the security threats, the most challenging task is to confirm the identification of the malicious node. The route discovery process is usually targeted during malicious routing intrusion, which leads to consequences where the mobile nodes don't comply with the assigned routing protocol [6]. Hence, it is essential to incorporate an effective routing scheme. Any mobile node that doesn't cooperate in data packet forwarding can be termed malicious and selfish nodes. However, there are more possibilities for defining the roles of such nodes. A regular node and malicious nodes are easier to identify; the challenge is only towards identifying any node with vulnerable behavior regarding routing. There is no statutory value to determine this scale of vulnerability to be confirmed as attacker node, compromised node, or selfish node. Various studies are being carried out to identify the malicious behavior in MANET [7]-[9]. However, there is still limitation associated with it, viz. i) excessive usage of encryption or complex secure routing techniques leads to loss of balance between communication and security, ii) adherence to a specific form of the attacker, iii) less emphasis towards dynamic attacking strategy. If this problem is not solved that the secondary issues owing to the presence of malicious nodes cannot be solved, viz. i) degradation in network connectivity, ii) isolated nodes with prominent declination in network performance, iii) less conservation of resources, iv) leads to exposure towards other forms of attackers. Hence, it is necessary to develop a security solution considering dynamic attackers without predefined information. In this perspective, the proposed system considers using Dynamic Source Routing (DSR) protocol owing to its beneficial features of being a reactive protocol in MANET [10].

However, there is less number of studies being investigated towards harnessing the potential of DSR protocol as the baseline of secure routing. It is also associated with a limitation which doesn't facilitate the mobile nodes to carry out operations towards identifying malicious behavior of mobile nodes. The proposed system uses DSR and improves its functionalities towards incorporating neighborhood monitoring states of the mobile nodes. The novelty is to detect and prevent unknown attackers of dynamic form present in MANET, whereas existing approaches mainly deal with the static form of attackers. Another novelty is about retaliation node to perform outlier management to present a unique mitigation measure. The paper's organization is as follows: Section II discusses existing approaches, while its limitation in the form of the research problem is addressed in Section III. The proposed methodology is discussed in Section IV. System

design is elaborated in Section V. Section VI discusses the outcomes obtained, while discussion of learning outcomes is presented in Section VII. The conclusion is given in Section VIII.

II. Existing Techniques

This section discusses the existing approaches towards securing Wireless Adhoc Network (WANET) towards identifying and mitigating malicious behavior of nodes. Current methods have witnessed the usage of trust-based mechanisms for mitigating malicious behavior. The work carried out by Abbasi et al. [11] has developed a scheme to secure message exchange operation towards vehicular Adhoc networks using a clustering algorithm concerning vehicle reputation. Irrespective of sound validation of the model, this approach doesn't resist the threat if the attacker is unknown. Another trust-based method was formulated by Chen et al. [12], considering a case study of internet-of-vehicles. The study has presented a collaborative scheme for filtering the behavior of regular to malicious nodes based on small-time intervals. The system also facilitates indirect trust calculation based on obtained recommendations from the adjacent nodes. The study is limited to resisting collusion attacks. Hence, the reputation factor also plays a significant role in identifying the malicious nature of nodes in WANET. A study towards emphasizing reputation is carried out by Guaya-Delgado et al. [13], where a unique routing scheme is constructed based on source routing. The core idea of this study is to identify any form of abnormal changes in the behavior of nodes considering routing behavior to be stationary. The study's outcome is limited for its effectiveness only concerning resisting selfish nodes in MANET.

Further study towards trust-based communication is presented by Dhananjayan and Subbaiah [14], where the conventional AODV protocol is rendered more secure considering all the essential constraints in MANET. The prime notion of this study is to match the ID of the packet sequence obtained from log traces of adjacent mobile nodes to assess the rate of trust. The target is to resist the generation of any form of a report by the malicious node. The study inherits the limitation as stale routing issues in AODV are not addressed. Janani and Manikandan [15] have presented a trust management scheme using Evidence Theory and Bayesian Theoryto validate malicious behavior detection. The study has emphasized understanding the degree of uncertainty in determining malicious behavior. However, the study offers a highly iterative scheme to identify malicious behavior inapplicable towards dynamic attackers in the MANET environment. A study toward evidence theory is further carried out by Mowla et al. [16]. The idea was to develop a cognitive learning scheme for the detection and mitigation of the jamming attack. The assessment of the study is carried out using a standard attack dataset for jamming intrusion. The applicability of the study is limited to jamming attacks only.

Further improvement in trust-based schemes towards malicious behavior detection is reported in Khan et al. [17], where multi-trust attributes are considered over-optimized link-state routing. The study is meant to resist multiple security threats; however, the formal model verification doesn't consider the dynamic alteration of attack strategy in MANET. The study of Kavitha et al. [18] has presented a security technique using optimized features followed by a classification scheme in MANET. The idea is to safeguard the Adhoc network from isolation attacks where neural network and particle swarm optimization have been used to optimize the features. Similar machine learning is used for classifying the attacker node. One of the potential pitfalls of this study is its dependency on training operations, which may bypass sure attackers using different strategies. Faisal et al. [19] have carried out a study to resist replication attacks, Sybil attacks, and impersonation attacks based on their received signal strength. This is done to find out the usage of additional hardware by the attacker. The security effectiveness of this study is limited to only the attacks mentioned above.

At present, blockchain was also reported to be used for resisting threats in MANET concerning its malicious behavior, as reported in the work of Ran et al. [20]. The researcher has used the AODV protocol, where blockchain is used for network development considering constraints. Although this is quite a novel approach with more applicability on future network technologies, the deployment of blockchain constructed is higher centralized, affecting the scalable performance in MANET. A unique study is carried out by Yasin and Zant [21], where the authors have presented a bait technique using a timer to identify and isolate attacks; however, the study applicability is limited to resist blackhole attacks only. Wireless Sensor Network is also a form of WANET system where existing approaches have been witnessed to mitigate identification issues of threats. The work of Alghamdi et al. [22] has used a convolution technique that generates security bits to resist attacks from malicious nodes using convolution codes. The simulated outcome shows the proposed scheme to offer better overhead control with better data transmission performance. However, the model doesn't provide prevention from any form of internal attacker. Wang et al. [23] have carried out a study where the bio-inspired protocol is designed for trust evaluation. The study offers the benefits of optimization, but it is not meant for resisting dynamic attackers.

Recent studies are being carried out where Dynamic Source Routing (DSR) is seen as the better option for securing threats in MANET. There is the evolution of various security-based schemes on DSR protocol viz. Bio-inspired algorithm (Almazok and Bilgehan [24]), cognitive-based DSR (Begum et al. [25]), trust-based scheme (Ishmanov and Zikria [26]), distributed key-generation (Kojima et al. [27]), path reliability-based approach (Liang et al. [28]), resisting blackhole attack (Mohanpriya et al. [29]), digital signature-based acknowledgment scheme (Srivastava et al.[30]).

From the above discussion of existing studies, it can be seen that there are good availability of research work towards securing wireless adhoc network. The methods mentioned above use different mitigation measures considering a specific case study in an Adhoc environment. Irrespective of reported benefits, there are various pitfalls towards the mechanism for securing adhoc network. Apart from this, it is also seen that existing approaches are deployed towards singular form of adversary whereas it is really a challenging one to identify if

the attacker alters its behaviour in WANET environment. Therefore, there is a need of a study which can address such issues. However, these schemes are more addressed towards specifics of adversary environment in deployment scene. The following section presents possibilities of various pitfalls associated with existing security approaches.

III. Research Problem

After reviewing the existing security approaches from the prior section, the following conclusive remarks has been drawn associated with the limitation of the methods:

- Scattered Approaches for Malicious Behaviour: There are split approaches towards different variants of the Adhoc network. However, there is no generalized scheme that is meant for all the variants of the Adhoc network. Due to varying operations in sensor networks and mobile Adhoc networks, a particular method cannot be used to secure both. Eventually, malicious behavior of a single form of attack will implicate different forms of consequences in other variants of the Adhoc network. There is less availability of standard generalized scheme towards, which impose as an impediment towards security incorporation.
- Unproportionate Schemes of Implementation: It has been seen that trust-based schemes are frequently used to resist attacks. However, trust-based schemes are developed based on direct and indirect trust computation without considering the dynamicity involved in the topology in a mobility environment. Apart from this, not much emphasis is offered to support a decentralized environment in a mobility environment. Other schemes are also available, e.g., evidence theory, machine learning, encryption, etc. However, their systems are still evolving, and no significant benchmarking is carried out to prove their applicability and effectiveness.
- Biased adversary model: Most studies have predefined information of the adversary, where it is not a challenging task to stop such an attack. However, these schemes are not applicable if the adversary changes its attack strategies. None of the existing studies are reported to work on dynamic attackers, which render the non-applicability of such algorithms in large-scale environments.
- More emphasis on prevention: A better form of prevention requires a better aggregation of the adversary's attack strategies. However, as the existing approaches have apriori information about the attackers, such prevention is only ensured for considered attacks and not for unknown attackers.
- More miniature Simplified Modelling: In the case of MANET, the mobile nodes consistently drain their energy and other resources. Hence, a robust security approach that requires consistent neighborhood monitoring will drain more energy for such resource-constrained mobile nodes. This demands much lightweight security operation, whereas the existing scheme offers more complex forms of the process that are less practical in the real world.

Hence, based on the research mentioned above problem, the statement is "identification of malicious behavior for the unknown attacker in decentralized mobility environment in Adhoc network is challenging task". The following section outlines the adopted research methodology which is meant for addressing the above mentioned research problems.

IV. Research Methodology

The implementation of the proposed system is carried out as an extension of our prior model of retaliation to identify the selfish node. This part of the implementation contributes to i) developing a novel secure DSR protocol and ii) the unique identification and prevention strategy towards malicious behavior of an unknown attacker in the MANET environment. The implemented architecture of the proposed system is as follow:

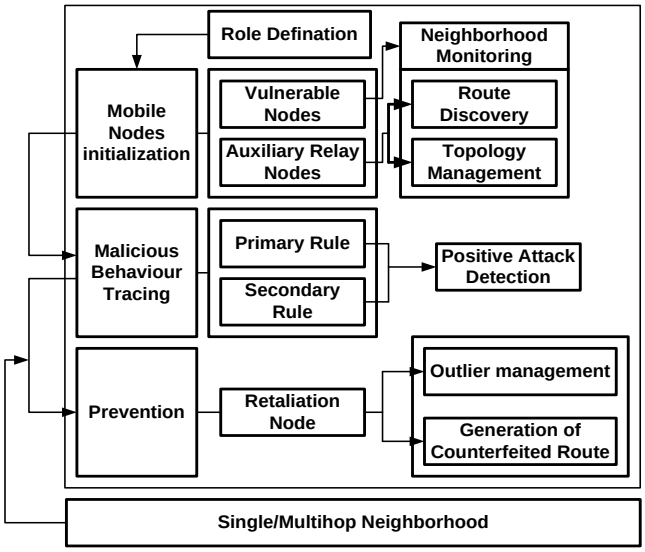


Fig. 1. Proposed Architecture of Secured DSR

Fig. 1 highlights the top-down architecture of the proposed secured DSR protocol in MANET. There are three significant operation blocks, i.e., i) mobile node initialization, ii) malicious behavior tracing, and iii) prevention. The first block of operation incorporates novel roles in mobile nodes by introducing vulnerable nodes and auxiliary nodes, unlike any existing approach of using DSR in MANET. The second block of process traces malicious behavior where two significant rules are developed to confirm the vulnerability as positive attack nodes. This rule formation assists in performing scrutiny of control messages of all nodes to ascertain their legitimacy using a novel algorithm for detecting a threat. Once the threat is positively detected, the last block of operation prevents the attackers by introducing a retaliation node. The novelty of the proposed system is the mechanism of resisting threats by retaliation node revised from our previous study. This node doesn't have its physical presence, and it is advertised by the target node under security observation when witnessed with the positive threat. The prime responsibility of this node is first to assess the update information of single and multihop

neighborhood nodes of vulnerable nodes. It generates counterfeited routes that don't match with any of the routes maintained in hop tables. To perform accurate detection, a probability-based computation is carried out to find out outliers in the detection process. The novelty of this approach is that it deploys a mechanism without using conventional encryption to secure the complete network system. It also introduces a special form of retaliation node which carries out this process of security incorporation. The following section discusses the system design.

V. System Design

This section discusses the proposed scheme implementation targeted towards mitigating routing misbehavior in DSR protocol in MANET. This section discusses the strategies used for designing the scheme and algorithm implementation.

A. Strategies Towards DSR Implementation

DSR is an on-demand routing scheme that utilizes a source-routing scheme that specifies the data sender's routes (complete or partial). The address of all the mobile nodes is required to determine the source route while performing route discovery in MANET. Caching is carried out for the aggregated information of path by the mobile nodes, which are used for the forwarding data packet. As the routed data consists of participating mobile nodes; hence, there are chances of intrusion and high overhead for large-scale network topology. Although conventional DSR protocols offer hop-by-hop communication for packet transmission to resist this overhead issue, identifying the unknown threat and disclosing the information is the biggest challenge. The mechanism of DSR is shown in Fig. 2 that carries out the a) route discovery and b) updating process in the MANET environment.

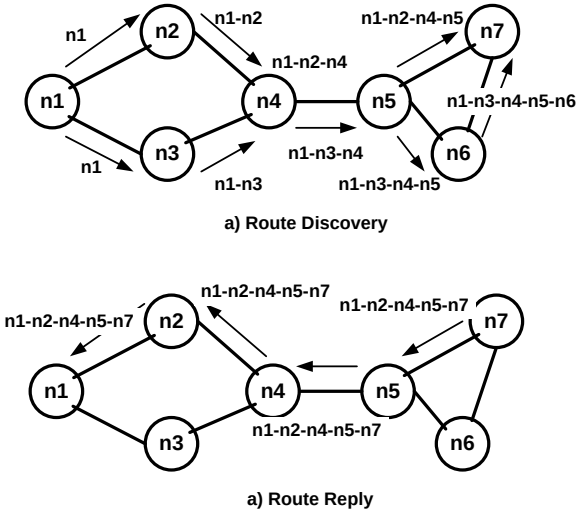


Fig. 2. Mechanism of Conventional DSR Scheme.

Implementing DSR protocol requires formulating two routing stages: i) discovery of routes and ii) route maintenance. The generation of route response beacon in DSR is carried out only when the destination mobile node has successfully. The destination node should possess the route information towards the original mobile node to forward the response beacon in MANET using DSR. Apart from this, the prominent dependency in DSR is to ensure all symmetric links find the route based on the header information. In case of faulty transmission, DSR protocol initiates maintenance, allowing the mobile node to generate error routes packets. This eliminates faulty links from the cached information, resulting in truncating all the faulty links generated from that mobile node. The updated routes are explored using the route discovery stage. Hence, to formulate a robust, secure strategy, it is necessary to realize the strength and weaknesses of the conventional DSR scheme in MANET. Realization of weakness will better formulate a proposed security solution that balances communication and security performance. The beneficial aspects of the DSR protocol are its adoption of an on-demand scheme that doesn't have any dependency on the regular transmission of update messages not to flood the network. The construction of the routes in DSR is carried out only when required, allowing the security system to monitor the threat closely. DSR significantly reduces the control overhead by ensuring utilization of route cache information by the relay node in MANET, which will be an added advantage towards supporting security operation. However, there are various discrepancies towards deploying conventional DSR in its actual form in security. The primary security challenge in DSR is that the broken link is not locally repaired while performing maintenance tasks, which attracts intrusion towards this broken link. The secondary security challenge in DSR is the prevalence of stale cached information of the route that results in critical inconsistencies while formulating new routes. This problem will render the attacker using the close information and intruding on the network during the route reconstruction phase. The tertiary security challenge in DSR is that it has good supportability for the environment with the static and lower extent of mobility; it is not eventually meant to resist security threats in increasing mobility environment. This could also result in potential delay and routing overhead. Hence, the proposed scheme assists in overcoming this pitfall of conventional DSR to mitigate a higher degree of threats in the MANET environment.

In the proposed scheme exhibited in Fig. 3, a novel mechanism is designed that lets all the mobile nodes carry out communication and identification of threats followed by resisting them simultaneously. The novelty of this scheme is to resist attacker nodes by deviating them in counterfeited routes which are applicable for multiple attackers of dynamic form in MANET. Unlike existing security approaches, the proposed system of DSR doesn't use any form of encryption and yet offers better protection with conservation of resources and efforts of mobile nodes towards executing security operations. This model provides counterfeited route information as the mitigation solution towards resisting attackers in MANET.

B. Proposed Enhanced DSR Implementation

The proposed system incorporates a certain degree of novelty to address the three essential loopholes discussed in the prior section in the DSR protocol. This is carried out towards strengthening DSR protocol for making it a high potential for identifying and mitigating threats. Following are the set of novel features introduced in the proposed secure DSR implementation in the MANET environment:

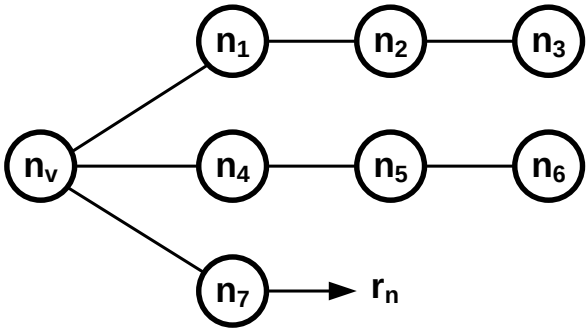


Fig. 3. Mechanism of Proposed DSR Scheme.

- The proposed model doesn't depend on any trusted third party or any other centralized scheme using DSR protocol.
- Unlike the existing DSR protocol, the proposed secured DSR protocol will assess the integrity by computing the inconsistency between the defined MANET topology and HELLO message.
- The proposed study appoints a specialized mobile node called an auxiliary relay node responsible for the broadcasting HELLO message and has the privilege to alter the topology to ensure faster route-finding towards the destination. Hence, the idea is also to protect this auxiliary relay node from threats.
- If there is no inconsistency, the model selects a single auxiliary relay node. However, in case of contradiction, the auxiliary relay node can be elected for all double-hop neighboring nodes where the auxiliary node plays the role of the single access point. However, neighboring nodes in double-hop with other paths are not privileged to select as a particular auxiliary relay node.

The properties mentioned above are incorporated within the DSR protocol. The prime justification of the above points is as follows: in case of any unknown attack, the initiation point will always be from the topology control message introduced in the proposed system. Eventually, it is impossible to stop the attacker (or vulnerable mobile node) from propagating legitimate control messages for topology control. Suppose the attacker broadcasts a counterfeited control message that eventually declares his presence, leading other mobile nodes to safeguard themselves. Therefore, a counterfeited control message is potentially a prominent strategy for the attacker. It can isolate the vulnerable node from the network entirely without the vulnerable node knowing about it. So, the idea of the proposed secure DSR protocol is to present a solution towards misleading the attacker node by using a retaliation node. In this case, the retaliation node will present a series of counterfeited information of node address to attacker or victim node once identified. The attacker/victim node will comply with the proposed protocol and will not deny the list of routes they are supposed to traverse. This will lead to the attacker node utilizing all its resources to find the counterfeited nodes, and this process will continue until they expend all its resources. After delivering the address list to attackers, the retaliation node is eliminated, and hence the network is fully secured even if the attacker changes its strategy. The following section discusses algorithm implementation as a part of the proposed solution.

C. Algorithm Design

This algorithm is responsible for determining the unknown threat present in the MANET environment. The complete discussion of the algorithm is carried out, referring to Fig. 3. The parameters used by the algorithms are i) mobile nodes V, ii) vulnerable node nv, which are part of normal mobile nodes V, iii) retaliation node rn that is advertised by node n7, iv) single and multihop neighboring nodes, i.e., sh(nv) and mh(nv) respectively, v) primary auxiliary relay node aux(nv) that maintains series of single-hop nodes of nv who elected nv as their auxiliary relay node, vi) secondary auxiliary relay node aux_sh(nv) which is a series of single-hop mobile nodes that are elected by the nv as the primary auxiliary node. The steps of the algorithm are as follows:

Algorithm for Determining the Threat

Input: *N* (mobile nodes)
Output: *b* (determining the state of threat)
Start
1. **For** i=1:N
2. $n_7 \rightarrow f_1$("HELLO||sh(n_v,n_2,n_5,r_n)")
3. n_v confirms (f_1(HELLO)||exclude(sh(n_v)))
4. **For** j=1:k
5. **If** Δ=1
6. **For** Δ≠g(HELLO)
7. **For** Δ≥dist(mh(n_v))
8. **If** n7→elect[λ(sh(n_v))]
9. nv checks Δ
10. n7 elects Δ
11. else
12. b=Δ elects n7 as aux_node
13. **End**
14. **End**
15. **End**
16. **End**
End

This algorithm assesses the control message disseminated by the mobile node in MANET using the proposed secured DSR protocol. Referring to Fig. 3, the algorithm considers single and multihop nodes of vulnerable mobile node nv as,

$$sh(n_v)=\{n_1, n_4, n_7\}$$

$$mh(n_v)=\{n_2, n_5\} \quad (1)$$

According to the proposed system, the vulnerable node nv should elect a secondary auxiliary relay node aux_sh(nv) which is a matrix retaining information about first-hop node n1 and n4; this is done by proposed secured DSR so that it can protect the request propagation to multihop nodes, i.e., mh(nv) which retains information about neighboring nodes of nv. Therefore, in the presence of the unknown adversary, the mobile node n7 will advertise a counterfeited control message which consists of single-hop information, i.e., sh(nv) consists

of vulnerable node nv and multihop nodes (n2, n5) and node n7(Line-1). However, the proposed system will not allow announcingn1 and n2 as it is possible for nv to validate this by considering the control message (HELLO) of node n7 with the control message of node n1 and n4. So, the proposed algorithm set up a protocol which is as follows: the vulnerable node nv should confirm that the announced mobile node by it should not match with the elements of matrix storing single-hop node information of vulnerable node sh(nv) (Line-3). This operation is carried out during the announcement of the control message by node n7, which possesses information about single-hop node information of n7.

On the other hand, node n7 should opt for an auxiliary relay node that will permit the usage of n1 and n4 (as they will be present in matrix storing mh(n7)). As a security mechanism, node n7 will falsely act as it opts for vulnerable node n7 as an auxiliary relay node to encapsulate its single-hop nodes n1 and n4. According to the proposed algorithm, the vulnerable node nv cannot infer that node n7 is an attacker. On the other hand, the vulnerable node nv can assess if the node n7 selects some of the auxiliary relay nodes for encapsulating multihop nodes of n7, i.e., mh(n7) with single-hop nodes of vulnerable node, i.e.,n1 and n4, i.e., It will eventually mean to assess the target node as n2 and n5 to be protected. Hence, the proposed algorithm set up another protocol that states: For a target k mobile node included within a control message (Line-4), the vulnerable node nv assesses the presence of another node Δ which are the subset of single-hop neighboring nodes of k (Line-5). The unit value in Line-5 represents the binary condition of its presence. Apart from this, the algorithm also ensures that Δ is not present in the sender's control message using a search function g(x) (Line-6). It also checks if Δ is positioned at a distance dist of multiple hops (at least three hops) (Line-7). The study analyzes another assessment to match this condition to determine if node n7 has elected λ mobile node (Line-8), which is also a part of single-hop neighboring nodes of n7 as the auxiliary node for encapsulating Δ (Line-9). This is followed by the election of Δ by node n7 (Line-10). Otherwise, the algorithm lets Δ elect n7 as the auxiliary node (Line-12). The above-stated operation can be carried out by searching within the routing table consisting of topology control information. Apart from this, it is feasible for the attacker to bypass this security scan by announcing that its position is in single hope from all the nodes N. Hence, to mitigate this challenge, the proposed system considers vulnerable node nv to possess conflicting and susceptible control messages consisting of all single-hop neighboring nodes of it. The algorithm mentioned above is meant to execute the mentioned rules sequentially to determine the state of intrusion, determined by variable b (Line-12). In case of any conflict, the vulnerable node nv elects n7 as the only auxiliary relay node exclusively for the mobile nodes that are announced the control message of node n7. Hence, the proposed algorithm without including any conventional encryption approach or complex iterative approach, the proposed algorithm can trace out the potential malicious behavior of an unknown attacker in MANET.

The next part of the algorithm implementation is towards resisting intrusion in MANET using the proposed secured DSR protocol. In this algorithm, the retaliation node is designed, responsible for forwarding counterfeited node information to the attacker once the attacker is positively identified in the prior algorithm. The steps of the algorithm are as follows:

Algorithm for resisting attack

Input: r_n (retaliation node)
Output: C (forwarding counterfeited route)
Start
1. $\Delta \rightarrow$add (r_n) for rn$\notin$sh(nv)
2. **If** Step-1=True
3. Δ broadcast r_n and compute aux_{prob}
4. **If** aux_{prob}=false
5. remove r_n
6. **End**
7. Compute outlier
8. C: $r_n \rightarrow f_2$(attacker)

End

The retaliation node is defined by Δ, which doesn't exist in the real-time MANET environment to mislead the attacker. In this case, the node Δ adds a retaliation node such that they don't belong to a class of single-hop neighboring nodes of vulnerable node nv (Line-1). All the new nodes Δ announce and advertise the information of the retaliation node if the first step is valid (Line-2). Further, it also computes the probability of auxiliary relay node (Line-3) considering all the mobile nodes concerning single-hop neighboring nodes of vulnerable node nv. When the probability computation for the auxiliary relay node is false (Line-4), the retaliation node rn is removed (Line-5). There is a benefit for this step, as after the retaliation node offers information of counterfeited nodes that don't exist to the intruder, the retaliation node must be eliminated. This will remove all the possibilities of intruders attempting to understand the strategy of proposed mitigation measures.

Further, the algorithm will compute outliers, calculated by calculating a total number of vulnerable nodes positively identified divided by a single hop neighborhood of vulnerable node nv (Line-7). In this case, the probability is computed for arbitrarily selected mobile nodes that are wrongly designated as an attacker, leading to vulnerable node nv preventing it as its auxiliary relay node. Hence, the system computes the possible distrusted mobile node to be part of vulnerable node nv, and it is a subset of single-hop neighboring nodes of nv. Then, the algorithm formulates a data transmission function f2(x) that advertises some randomly selected nodes that are not the common node to the attacker (Line-8). Upon receiving this information, the attacker will need to trust the data and select the routing towards the counterfeited path C leading to complete energy drainage. This mechanism doesn't lead to any form of halt in the communication system of regular nodes. The novelty of this algorithm is that it doesn't perform any form of iterative operation and offer a dual-layer of checks to confirm the malicious behavior of an unknown mobile node. The following section discusses simulation outcomes obtained by implementing the proposed algorithm.

VI. Result Analysis

The implementation of the proposed logic is scripted in MATLAB. The analysis considers 500-1000 mobile nodes deployed within 1000 x 1000 m2 simulation area. The assessment is carried out for 1000 simulation rounds where the test environment is created for node density while four performance parameters are chosen to be evaluated viz. i) proportion of retaliation node, ii) Throughput, iii) proportion of required Auxiliary Relay Node (ARN), and iv) processing time. The comparison is carried out concerning conventional DSR protocol. The prime justification is that the proposed system offers security by incorporating a series of DSR protocol changes without including any conventional encryption or security system in MANET. This makes it more suitable to be compared with traditional DSR protocol, which lacks any security incorporations.

A. The Proportion of Retaliation Node

The computation of the retaliation node is carried out by consistently monitoring the available number of retaliation modern while implementing the algorithm. Cost effective security solution will anticipate reduced dependencies of retaliation node, which is analyzed in this part of result analysis. The observation is carried out over increasing node density, representing increasing traffic load and an unknown attacker.

Fig. 4 highlights that the proposed system progressively reduces the retaliation node's dependencies, which can be justified as follows: According to the algorithm, the retaliation node doesn't exist in real-time. It only exists in the form of memory which forwards counterfeited information to attackers. Once the attackers comply with its generated path of rn, these memories are disposed of off completely. In case of further attack, the vulnerable node is now aware of the attacker node identity. Hence, without even using the retaliation node, the vulnerable node can isolate themselves from any form of communication (route discovery and data forwarding) and therefore be safe without depending on the retaliation node. Hence, the dependency of rn decreases with the increase of traffic.

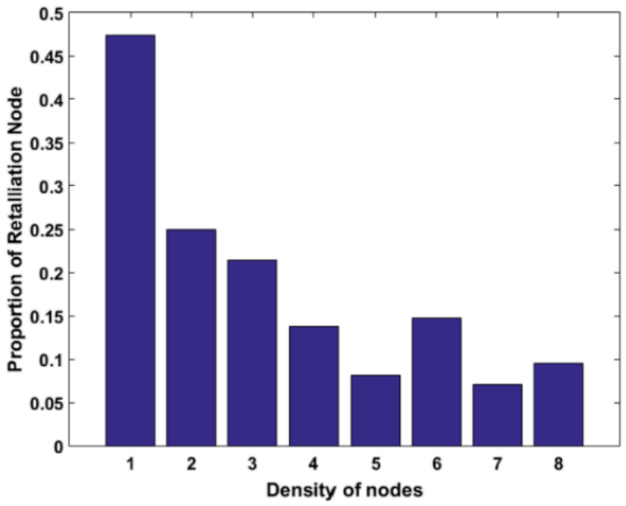


Fig. 4. Proportion of Retaliation Node.

B. Throughput Analysis

The proposed system computes throughput by monitoring the cumulative data received by the destination mobile nodes in the MANET environment measured in kilobytes per second.

The assessment is carried out separately for both DSR and Secured DSR protocol (proposed) in the presence of increasing node density. The outcome exhibited in Fig. 5 highlights that the proposed system offers better throughput than the existing system. The major reason is that conventional DSR protocol suffers from the challenge of retaining stale route information, which causes usage of similar routes for a specific range of data transmission attempts iteratively. In the presence of an attacker, this route is often compromised, leading to lower availability of channel capacity, causing degradation in throughput. On the other hand, the proposed system maintains the parallel process of identification and prevention using retaliation node without any effect on ongoing communication. This causes the proposed method to be less sensitive towards an attacker's presence once they are positively identified.

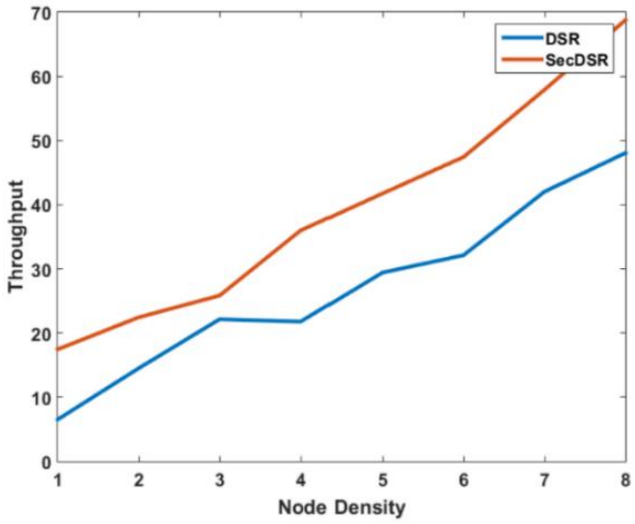


Fig. 5. Proportion of Throughput.

C. The Proportion of Required ARN

Auxiliary Relay Node (ARN) plays a contributory role in the proposed study. It holds the responsibility of route discovery for data transmission and topology management in case of the attacker's presence in the proposed DSR protocol. However, it should also be noted that there is no physical presence of a retaliation node in the proposed system. In contrast, ARN has a physical presence in the specially selected node based on the highest resource contents. All work carried out in MANET only uses mobile nodes, whereas the proposed system uses mobile nodes and ARN as a novelty. So it is wise enough to find out if such an ARN node affects the network with an increasing traffic load. For the system to be more secure, it is required to reduce the number of ARN with increasing traffic as the privilege of ARN to control topology is a prime attempt to be compromised by the attacker. Fig. 6 highlights that the proposed system offers significantly fewer ARN nodes in comparison to the existing DSR. The DSR protocol does not include the ARN concept; however, to

maintain a similar testbed, analysis is carried out to find the impact of the presence of ARN in regular DSR operation and proposed secured DSR operation. The calculation is carried out by monitoring a total number of instantaneous selected ARN during each iteration corresponding to node density. The prime justification behind this outcome is that: conventional DSR protocol uses source routing where there is a dependency toward retaining node address within the routed packet.

Further, the cached information of DSR keeps on increasing with increasing node density. So, technically DSR protocol doesn't offer much supportability towards the proposed usage of ARN. However, due to the formulated operation of the proposed algorithm with ARN, conventional DSR is found to be better when ARN is used. The proposed system further exceeds this outcome as ARN is selected only when a significantly new topology of MANET is encountered. The proposed method also keeps on computing outliers, ensuring that the detection rate is always higher, reducing dependencies on ARN.

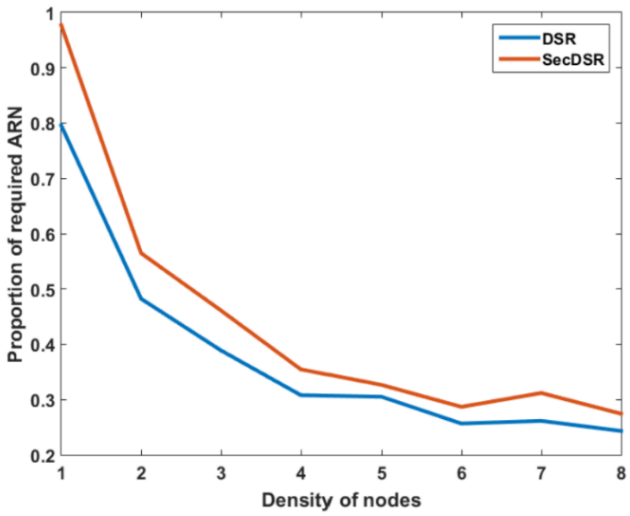


Fig. 6. Proportion of Required ARN.

D. Processing Time

Cumulative time required to execute the algorithm gives processing time, as exhibited in Table I.

TABLE I. ANALYSIS OF PROCESSING TIME

	DSR	SecDSR
Processing Time	8.825 s	1.271 s

The prime reason behind reduced processing time for the proposed system (SecDSR) is because it doesn't have an inclusion of much iterative operation concerning route discovery, maintenance, intrusion detection, and prevention as compared to existing DSR where higher iterative and more operative steps in involved to find the effective routes of communication in MANET environment. This causes reduced time for the proposed system in contrast to conventional DSR.

VII. DISCUSSION

From the prior section, it has been seen that proposed system carries out better performance outcome in contrast to existing secured DSR version. Following are the learning outcomes extracted from this model implementation:

- The proposed scheme is highly suitable for application working on WANET which is basically larger in dimension as well as distributed. This is because of the memory sharing concept by each nodes where updating the threat flags are quite faster compared to conventional DSR. Therefore, a better processing time is obtained.
- The proposed system ensures its applicability of dynamic attacker as well as multi-attacker at same time. It is because it can formulate an assessment with respect to unit hop link associated with the legitimacy of the attacker node. Hence, any node changes its strategy at any point of time will be instantly notify the other nodes about this chance. This causes faster identification of attack environment.
- The throughput of proposed system is quite good and this is because of the non-inclusion of iterative checks or sophisticated classes of operation. Along with assessing the legitimacy of the nodes, the prime target node can always perform seamless propagation of data to its destination node. Hence, data propagation is not affected by its assessment process towards intrusion.

Therefore, on the above ground of constructive outcome, it can be stated that proposed system offers a simplified and cost efficient computational solution towards securing WANET from potential threats.

VIII. CONCLUSION

The proposed system approaches resisting unknown attackers in MANET based on computation being carried out on malicious behavior. In the adversary, the study considers that the vulnerable node is manipulated by the attacker while playing the role of the auxiliary relay node. By doing this, the adversary will be able to gain access to the complete network. The proposed system offers a solution to resist it. The contribution of the proposed study is as follows:

- The proposed model assists in safeguarding the routes as well as nodes connected in MANET from being disclosed to the adversary without any form of dependencies of any trusted authority or third party.
- The proposed model introduced an auxiliary relay node which is responsible for performing route discovery as well as management of topology which reduces the same effort carried out by mobile nodes as seen in the existing system.
- The proposed method introduces a retaliation node which is a non-physical entity to mislead the attacker as a unique prevention strategy unlike any current mechanism in MANET.

The future work of the proposed system will be carried out further to optimize the study outcomes. Adoption of bio-inspired protocols can be adopted in order to find out more optimized solution towards delay and data propagation. More cases of multi-objective function can be formulated to ensure more resistivity towards physical attacks.

REFERENCES

[1] K. -H. Cho, S. -H. Lee and V. Y. F. Tan, "Throughput Scaling of Covert Communication Over Wireless Adhoc Networks," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7684-7701, Dec. 2020, doi: 10.1109/TIT.2020.3011895.

[2] B. Ojetunde, N. Shibata and J. Gao, "Secure Payment System Utilizing MANET for Disaster Areas," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 12, pp. 2651-2663, Dec. 2019, doi: 10.1109/TSMC.2017.2752203.

[3] X. Chen, T. Wu, G. Sun and H. Yu, "Software-Defined MANET Swarm for Mobile Monitoring in Hydropower Plants," IEEE Access, vol. 7, pp. 152243-152257, 2019, doi: 10.1109/ACCESS.2019.2948215.

[4] H.S. Bedi, S. Verma, M. Goel, "A Survey on MANET Security Challenges, Attacks and its Countermeasures", *International Journal of Advanced Research in Computer and Communication Engineering*, vol.5, Iss.8, 2016.

[5] K. Kumar, S. Verma, Kavita, NZ Jhanjhi, M N Talib, "A Survey of The Design and Security Mechanisms of The Wireless Networks and Mobile Ad-Hoc Networks", *OP Conf. Series: Materials Science and Engineering*, vol.993 2020.

[6] M. S. Khan, D. Midi, M. I. Khan, and E. Bertino, "Fine-Grained Analysis of Packet Loss in MANETs," *IEEE Access*, vol. 5, pp. 7798-7807, 2017, doi: 10.1109/ACCESS.2017.2694467.

[7] Motwani and Anand, "Survey of Malicious Attacks in MANET", *International Journal of Computer Applications*, vol.80, pp.28-30, 2013. Doi 10.5120/13931-1916.

[8] Bhatia, Tarunpreet & Verma, A., "Security Issues in Manet: A Survey on Attacks and Defense Mechanisms", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol.3, pp.1382-1394, 2013.

[9] F. Abdel-Fattah, K. A. Farhan, F. H. Al-Tarawneh and F. AlTamimi, "Security Challenges and Attacks in Dynamic Mobile Ad Hoc Networks MANETs," *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology*, pp. 28-33, 2019, doi: 10.1109/JEEIT.2019.8717449.

[10] G. Toso, R. Masiero, P. Casari, M. Komar, O. Kebkal and M. Zorzi, "Revisiting Source Routing for Underwater Networking: The SUN Protocol," IEEE Access, vol. 6, pp. 1525-1541, 2018, doi: 10.1109/ACCESS.2017.2779426.

[11] R. Abassi, A. B. C. Douss, & D. Sauveron, "TSME: a trust-based security scheme for message exchange in vehicular Ad hoc networks" *Springer Open-Human-centric Computing and Information Sciences*, Article No.43, 2020.

[12] J. Chen, T. Li, and J. Panneerselvam, "TMEC: A Trust Management Based on Evidence Combination on Attack-Resistant and Collaborative Internet of Vehicles," *IEEE Access*, vol. 7, pp. 148913-148922, 2019, doi: 10.1109/ACCESS.2018.2876153.

[13] L. Guaya-Delgado, E. Pallarès-Segarra, A. M. M. & J. Forné, "A novel dynamic reputation-based source routing protocol for mobile ad hoc networks", *Springer-EURASIP Journal on Wireless Communications and Networking*, Article No. 77, 2019.

[14] G. Dhananjayan & J. Subbiah, "T2AR: trust-aware ad-hoc routing protocol for MANET", Springer-Open, Vol.5, Article No.995, 2016

[15] J. V S & Manikandan M S K, "Efficient trust management with Bayesian-Evidence theorem to secure public-key infrastructure-based mobile ad hoc networks", *Springer-EURASIP Journal on Wireless Communications and Networking*, Article number: 25, 2018.

[16] N. I. Mowla, N. H. Tran, I. Doh and K. Chae, "Federated Learning-Based Cognitive Detection of Jamming Attack in Flying Ad-Hoc Network," *IEEE Access*, vol. 8, pp. 4338-4350, 2020, doi: 10.1109/ACCESS.2019.2962873.

[17] M. S. Khan, M. I. Khan, Saif-Ur-Rehman Malik, Osman Khalid, Mukhtar Azim & Nadeem Javaid, "MATF: a multi-attribute trust framework for MANETs", *Springer- EURASIP Journal on Wireless Communications and Networking*, Article number: 197,2016.

[18] T. Kavitha & K. Geetha & R. Muthaiah, "India: Intruder Node Detection and Isolation Action in Mobile Ad Hoc Networks Using Feature Optimization and Classification Approach", *Wiley-Journal of Medical Systems*, vol.43, Iss.179, 2019.

[19] M. Faisal, S. Abbas & H. Ur Rahman, "Identity attack detection system for 802.11-based ad hoc networks", *EURASIP Journal on Wireless Communications and Networking*, Article number: 128, 2018.

[20] C. Ran, S. Yan, L. Huang & L. Zhang, "An improved AODV routing security algorithm based on blockchain technology in ad hoc network", *EURASIP Journal on Wireless Communications and Networking*, Article number: 52, 2021.

[21] A. Yasin and. A. Zant, "Detecting and Isolating Blackhole Attacks in MANET Using Timer Based Baited Technique", *Wiley-Hindawi Wireless Communications and Mobile Computing*, 2018.

[22] T. A. Alghamdi, "Convolutional technique for enhancing security in wireless sensor networks against malicious nodes", *Human-centric Computing and Information Sciences*, vol.9, Article number: 38,2019.

[23] Y. Wang, M. Zhang & W. Shu, "An emerging intelligent optimization algorithm based on trust sensing model for wireless sensor networks" *EURASIP Journal on Wireless Communications and Networking*, Article number: 145, 2018.

[24] S. A. Almazok & B. Bilgehan, "A novel dynamic source routing (DSR) protocol based on minimum execution time scheduling and moth flame optimization (MET-MFO)", *EURASIP Journal on Wireless Communications and Networking*, vol.2020, Article number: 219, 2020.

[25] S. Begum, Y. Nianmin, S. B. H. Shah, A. Abdollahi, "Source Routing for Distributed Big Data-Based Cognitive Internet of Things (CIoT)", *Wiley-Hindawi Wireless Communications and Mobile Computing*, 2021.

[26] F. Ishmanov and Y. B. Zikria, "Trust Mechanisms to Secure Routing in Wireless Sensor Networks: Current State of the Research and Open Research Issues", *Hindawi-Journal of Sensors*, 2017.

[27] H. Kojima, N. Yanai, and J. P. Cruz, "ISDSR+: Improving the Security and Availability of Secure Routing Protocol," in IEEE Access, vol. 7, pp. 74849-74868, 2019, doi: 10.1109/ACCESS.2019.2916318.

[28] Q. Liang, T. Lin, F. Wu, F. Zhang, W. Xiong, "A dynamic source routing protocol based on path reliability and link monitoring repair", *PLOS ONE, Open Access*, 2021.

[29] M. Mohanapriya, N. Joshi, M. Soni, "Secure dynamic source routing protocol for defending blackhole attacks in mobile Ad hoc networks", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 21, No. 1, pp. 582~590, 2021.

[30] A. Srivastava, S. K. Gupta, M. Najim, N. Sahu, G. Aggarwal & B. D. Mazumdar, "DSSAM: digitally signed secure acknowledgement method for mobile ad hoc network", *EURASIP Journal on Wireless Communications and Networking*, vol.2021, Article number: 12, 2021.

Free Hardware based System for Air Quality and CO2 Monitoring

Cristhoper Alvarez-Mendoza[1], Jhon Vilchez-Lucana[2], Fernando Sierra-Liñan[3], Michael Cabanillas-Carbonell[4]
Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, Lima, Perú[1, 2]
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú[3]
Vicerrectorado de Investigación, Universidad Norbert Wiener, Lima, Perú[4]

***Abstract*—Due to the increase in air pollution, especially in Latin American countries of low and middle income, great environmental and health risks have been generated, highlighting that there is more pollution in closed environments. Given this problem, it has been proposed to develop a system based on free hardware for monitoring air quality and CO2, in order to reduce the levels of air pollution in a closed environment, improving the quality of life of people and contributing to the awareness of the damage caused to the environment by the hand of man himself. The system is based on V-Model, complemented with a ventilation prototype implemented with sensors and an application for its respective monitoring. The sample collected in the present investigation was non-probabilistic, derived from the reports of air indicators during 15 days with specific schedules of 9am, 1pm and 6pm. The results obtained indicated that the air quality decreased to 670 ppm, as well as the collection time decreased to 5 seconds and finally the presence of CO2 was reduced to 650 ppm after the implementation of the system, achieving to be within the standards recommended by the World Health Organization.**

***Keywords*—Air quality; air pollution; co2; control system; free hardware; v-model**

I. Introduction

Air pollution is a negative effect that has been aggravated over the years, the development of the population has generated new demands in homes, industries and public centers, bringing as a consequence the continuous pollution of the environment and multiple diseases [1]. One of the main causes of the increase in carbon dioxide (CO2) is the burning of fuel [2].

Recent studies affirm that there is a higher concentration of pollutant gases in closed environments than outdoors [3]. The World Health Organization (WHO) [4], states that, due to indoor air pollution, public health, especially that of children and the elderly, is being affected. Currently the rate of premature death of children amounts to 7 million, this is due to the concentration of gases such as CO2, which generates suffering from asthma, reduced growth and respiratory problems.

Air quality is expressed by the high concentration of several gases, the main one being CO2, likewise the WHO mentioned that the allowed value to improve air quality in indoor environments is between 800 and 1000 ppm (unit to measure the volume of particles), due to the pollutants found in the air [4]–[6]. According to the report presented by Air Quality Life Index (AQLI) in the last year [7], it was identified that half of the Latin American population receives dangerous levels of pollution, with Peru being one of the countries with the most critical points, having the inhabitants of its capital (Lima) with a life expectancy of 4.7 years reduced due to air pollution (Fig. 1).

Current technological and scientific progress has driven the development of systems that improve people's quality of life, providing welfare to the community by providing relevant and pertinent information for decision making. In the technological context of the Internet of Things (IoT), these systems involve the measurement and monitoring of various environmental variables [8], supported together with the use of free hardware, which has had a growing use being used in any type of project without the need to purchase a license [9]. The data obtained from the electronic devices will later prove to be important for raising public awareness of the level of air pollution present in our environment, which damages our health.

The proposed control system aims to provide an alternative system based on free hardware for air quality management in a closed facility, in order to reduce the pollutant gases in that place. This will be achieved by making use of different components such as sensors and communication devices, interacting with each other and obtaining the data of the gases under study, to later analyze if they exceed the established standards and act automatically, finally turn on the necessary components and ventilate the environment.

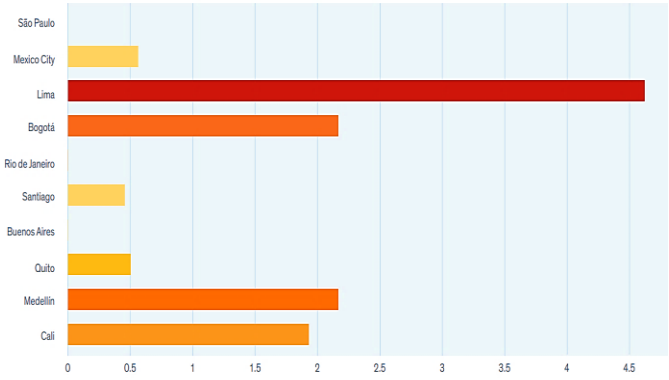


Fig. 1. Reduction of Life Expectancy Years due to PM2.5 Concentrations (Organic Chemicals Suspended in the Air) in 10 Largest Cities in Latin America [7].

This article has been organized in six different sections to achieve a better understanding: Section II contains the bibliographic review carried out after the study of previous research related to the subject under study. Section III details the development of the methodology applied, as well as the development of the proposed system. Section IV shows the results obtained after the pre- and post-implementation of the system, as well as the summary reports of each indicator evaluated. Section V contains the discussions, as well as the subsequent analysis of the results. Finally, section VI details the conclusions obtained from the development and implementation of the system.

II. Review of the Literature

The development of a control system is aimed at the systematic or manual management of the technological components in a simple and effective way, which is why we propose the development of a control system that helps reduce CO2 emissions by monitoring air quality to improve the purification and maintenance of this at normal levels.

According to what was investigated, it was possible to gather proposals of solutions from different authors regarding the problems that arise in the control of air quality, thus contributing to the research carried out. To achieve this, it was necessary to identify the degree of contamination and the quality of the air provided in the environment of a population and the factors that influence it [10], resulting to have a direct relationship with respiratory problems, cardiovascular, lung cancer, among other related diseases, this is due to the fact that the WHO guidelines [11], [12] were not taken into account. Regarding indoor air quality, the risk is doubled due to constant exposure to high levels of smoke [12], [13], therefore WHO has considered the training of countries with higher pollution rates through the use of monitoring technologies, workshops, etc.

Likewise, the contribution of the research [14] was considered, it proposed a system to measure the air quality in five routes by means of monitoring sensors in vehicles, collecting data from the cluster without the need for a fixed network structure, making recommendations in the course of travel through an application and social networks, thus contributing to the use of health-conscious transportation.

On the other hand, an IOT-based system was proposed through wireless networks based on connections through web services protocols, data transmission and data encryption because it is intended for smart buildings [15], proposing a free hardware design thinking about the functionality and scalability of it, aiming to predict and control the CO2 gas which is produced by people. The tools implemented proved to be efficient in recording data for future predictions and environmental controls, in addition to having great energy savings.

In Ref. [16], a 2-year study was conducted at different air quality stations based on the evaluation of particulate matter (PM10) in 3 districts of the province of La Oroya in Peru, one of the mining sites with the highest exposure to air pollution. The sources of emissions, the verification with respect to emission and immission standards and the type of pollutant that exists at the site were verified; by having this information, it was possible to determine the level of risk and environmental impact that existed at the site, evidencing an improvement over the years.

This section explores previous studies on the different tools that have been developed to control and monitor air quality in real time, which was necessary for the development of this research, identifying the sensors already used before. This is due to the need of people to know the state of air quality in their environment since it generally has a direct impact on health [17]. Table I summarizes recent developments in these tools and techniques.

TABLE I. Previous Studies in Real Time Air Quality Monitoring and Control

Tool	Ref.	Technology	Conclusions
AirCloud	[18]	Two air quality monitoring devices were developed using PPD42NJ sensors to acquire data.	A low-cost, personal air quality monitoring system capable of achieving good PM2.5 prediction accuracies was obtained.
WSN based AQM	[19]	Zigbee WSN was used, employing gas sensors for (CO2, NO2, ozone).	Clustering energy-efficient air sensor protocol was introduced to monitor air quality.
AirSense	[20]	Use was made of the MQ-135 sensor connected to the Arduino Pro Mini, the sensor data was collected using ThingSpeak.	A dual-purpose, low-power, low-weight system was obtained that was able to detect and collect data.
IoT-enabled AQM	[21],	Use was made of several sensors deployed on Marvel Board and AWS cloud.	It is less expensive and has high fidelity. Due to the cloud platform special limitations are overcome.
IoT-Raspberry PI	[22]	It required integrating different types of sensors with the Raspberry Pi, controlling and managing them all using Python.	Analyzes some hazardous gases such as carbon monoxide (CO), nitrogen dioxide (NO2) and other gases.
(IoT) based air quality monitoring system	[23]	It was able to address Malaysia's concern about the delay announcement and measurement of PM2.5 haze.	The system monitors common air pollutants such as particulate matter (PM) of PM2.5, PM10 and carbon monoxide (CO) gas.
Genetic algorithm and neural network	[24]	An optimized air quality estimation model was developed based on a genetic algorithm and an artificial neural network.	The optimized network model is used to estimate the carbon monoxide concentration in a polluted environment.

III. Methodology

The methodology used for the development of the project was the V-Model, which is used for software development by establishing the course of the project through phases, having as one of its main benefits to define the processes of quality management that can interact with the phases with each other

[25]. This model allows integration tests to be performed as the phases are completed, allowing to verify the progress of each of the add-ons, helping to avoid errors.

A. Population and Sample

The study population consisted of the number of reports of air quality indicators in the "24 de Junio" shopping center located in the city of Lima. The type of sample collected in this research was non-probabilistic, derived from the reports of air quality indicators during 15 days with specific hours of 9am, 1pm and 6pm.

B. Development of V-Model

1) Phase I – Specifications: In this phase all the functional and non-functional requirements for the construction of the hardware and also the application were conceptualized.

2) Phase II: Global Design: In this phase, the design of the overall solution was carried out. Fig. 2 shows the connection between the ESP8266 board and the MQ-135 (air quality) and MG-811 (CO2 sensor) sensors.

The complete system architecture for the subsequent air control operation and the application model is shown in Fig. 3.

The user visualizes the level of contamination present at the site, the board collects data from the sensors and sends it to both the application and the ventilation system to be turned on or off.

3) Phase III - Detailed Design: Fig. 4 shows the communication model of the devices used, where the Esp8266 Wi-Fi chip collects the information from the sensors and how it is sent to the application, and also shows the automation of the ventilation set. The user can visualize the air quality level of the shopping center from any place where he/she has an internet connection.

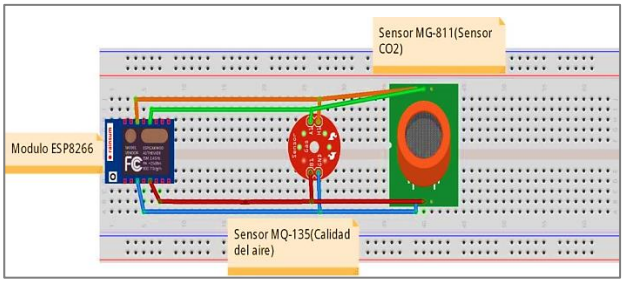


Fig. 2. Prototype Design.

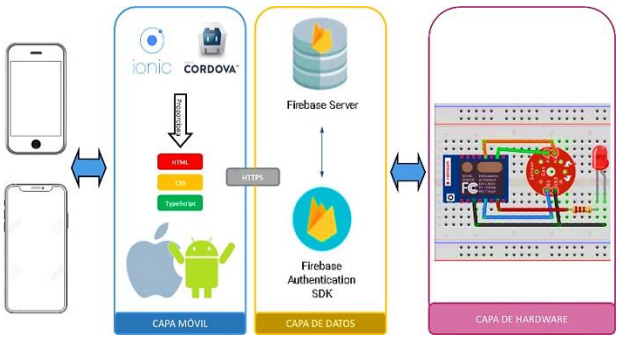


Fig. 3. General Architecture of the Proposed System.

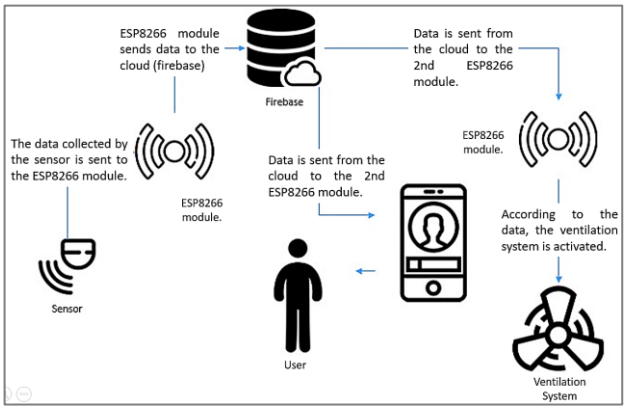


Fig. 4. Communication Model between Devices.

4) Phase IV – Implementation and Integration: In this phase, all the requirements of Phase I regarding the control system (Fig. 5) and the application (Fig. 7) were incorporated and implemented.

The system was programmed Fig. 6(a) and then the data captured by the measurement sensors were analyzed to verify the state of the air. After the sensor collected the data, they were analyzed internally in the ESP8266, first verifying if the air was within the allowed margins as shown in Fig. 6 (b).

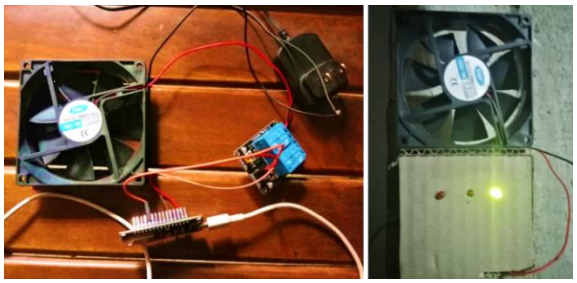


Fig. 5. Control System Integration.

```
StaticJsonBuffer<200> jsonBuffer;
JsonObject& root = jsonBuffer.createObject();
root["id"] = n++;
root["medicion"] = s_analogica_mq135;
root["medicionco2"] = co2ppm;
root["mensaje"] = mensaje;
root["ventilación"] = ventilacion;
root["zdate"] = Date;
// set value
Firebase.push("monitoring", root);
// handle error
if (Firebase.failed()) {
    Serial.print("setting /number failed:");
    Serial.println(Firebase.error());
    return;
}
```

(a)

```
COM3
Enviar
aire normal 445.00
aire normal 445.00
aire normal 442.00
aire normal 442.00
aire normal 442.00
aire normal 441.00
aire normal 444.00
aire normal 450.00
aire con nivel medio 451.00
aire normal 439.00
Autoscroll
Ambos NL & CR
9600 baudio
```

(b)

Fig. 6. Tests to the Developed Prototype.

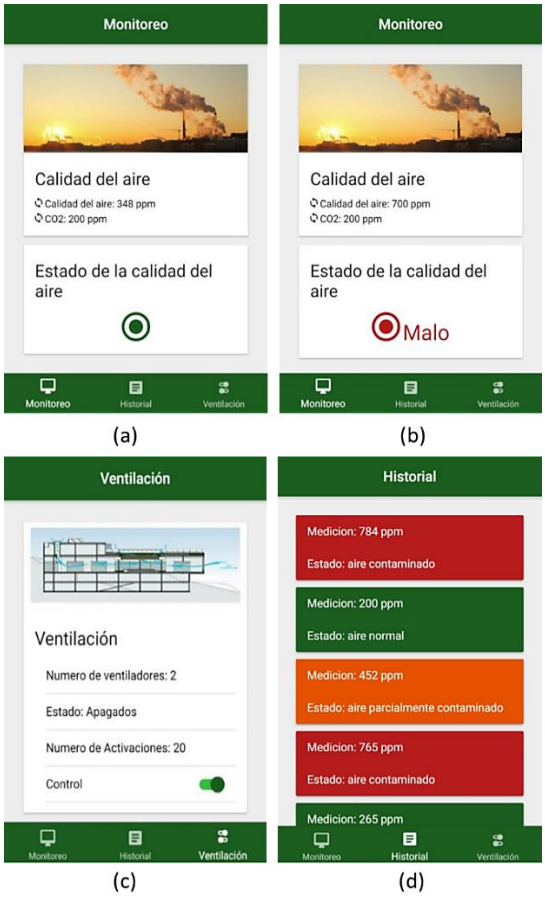


Fig. 7. Interfaces and Application Testing.

Fig. 7 shows the developed application which fulfills the function of receiving the data in real time, so that it can be seen by the users. In Fig. 7(a) and Fig. 7(b) the application shows the states (red, yellow and green) depending on the quality of the captured air, as specified in Table II. Fig. 7(c) shows that the application has a switch which allows the control of the ventilation system. In addition, in Fig. 7(d) the application shows the air history of the last 10 measurements of each hour specified in addition each shows a color depending on the air quality measurement in those ranges.

5) Phase V - System Operational Test: In the last phase, everything that was done was analyzed, drawing conclusions on the actions that went well and those that need to be improved in the project.

A pre-experimental design with pre- and post-test was applied to this research (1).

$$Ge \quad O_1 \quad X \quad O_2 \qquad (1)$$

TABLE II. STATES OF THE PROTOTYPE DEPENDING ON AIR QUALITY

States	Action
Good	The status turns green indicating that the air is clean.
Medium	The state turns a yellow color indicating that the air is not as pure.
Poor	The status turns red which indicates that the air is very polluted.

Where:

- Ge: Experimental group: the study is given to the air during 3 specific times in the day.
- O1: Pre-Test data for the dependent variable indicator before the implementation of the control system. Pre-Test measurement of the experimental group.
- O2: Post-Test data for the indicator of the dependent variable. Post-Test measurement of the experimental group.
- X: Control System = The object to be tested.

IV. RESULTS

Table III shows the data obtained in the Pre and Post Test of the KPI1, KPI2 and KPI3 of the present research.

TABLE III. RESULTS OBTAINED FROM THE PRE-TEST AND POST-TEST

# Days	Specific hours	KPI_1		KPI_2		KPI_3	
		Pre	Post	Pre	Post	Pre	Post
1	09:00 a. m.	950	700	790	660	300	6
	01:00 p. m.	1520	1270	1365	1120	350	5
	06:00 p. m.	1030	780	878	748	240	9
2	09:00 a. m.	870	620	716	586	290	7
	01:00 p. m.	1420	1170	1267	1100	300	7
	06:00 p. m.	930	680	771	641	240	5
3	09:00 a. m.	920	670	760	630	300	7
	01:00 p. m.	1260	1010	1095	920	270	8
	06:00 p. m.	870	620	712	582	250	5
4	09:00 a. m.	670	420	518	388	220	8
	01:00 p. m.	1326	1076	1172	1042	200	7
	06:00 p. m.	846	596	687	557	260	6
5	09:00 a. m.	720	470	556	426	260	6
	01:00 p. m.	1463	1213	1293	1163	240	7
	06:00 p. m.	823	573	656	526	240	9
6	09:00 a. m.	820	570	655	525	310	6
	01:00 p. m.	1260	1010	1093	963	305	6
	06:00 p. m.	680	430	525	395	230	6
7	09:00 a. m.	640	390	482	352	290	5
	01:00 p. m.	1050	920	885	800	270	6
	06:00 p. m.	820	570	665	535	233	9
8	09:00 a. m.	740	490	587	457	305	7
	01:00 p. m.	1160	910	990	900	240	5
	06:00 p. m.	901	651	737	607	120	6
9	09:00 a. m.	810	560	650	520	130	5
	01:00 p. m.	1362	1112	1201	1071	205	8
	06:00 p. m.	1025	775	867	737	100	5
10	09:00 a. m.	720	470	565	435	160	8
	01:00 p. m.	1060	810	904	774	200	6

# Days	Specific hours	KPI_1		KPI_2		KPI_3	
		Pre	Post	Pre	Post	Pre	Post
	06:00 p. m.	852	602	682	552	320	5
11	09:00 a. m.	860	610	705	575	203	5
	01:00 p. m.	1630	1380	1470	1340	156	5
	06:00 p. m.	945	695	784	654	198	5
12	09:00 a. m.	630	380	471	341	248	6
	01:00 p. m.	960	710	790	660	310	6
	06:00 p. m.	730	480	571	441	204	8
13	09:00 a. m.	790	540	600	470	302	7
	01:00 p. m.	1230	1000	1060	930	209	8
	06:00 p. m.	1040	790	870	740	281	8
14	09:00 a. m.	650	400	476	346	295	8
	01:00 p. m.	1020	870	842	712	333	7
	06:00 p. m.	720	470	568	438	220	6
15	09:00 a. m.	890	640	735	605	250	6
	01:00 p. m.	1530	1280	1367	1237	240	7
	06:00 p. m.	903	653	742	612	270	8

Table IV shows the average results obtained for each KPI of the research; these results were derived from the analysis of Table III.

The results in the three KPI's, both pre and post, performed 15 days later, show a decrease for each indicator (Table IV), which means that the control system was of great support to reduce air quality and CO2 pollution.

From Fig. 8, where each indicator is measured with different index tests which varies between 0 - 1200, it can be observed that the average of KPI I is the one that has been reduced the most with respect to the other KPI's, it can be said that there is a better air quality.

TABLE IV. RESEARCH INDICATORS

Indicator	Pre-Test	Post-Test
KPI I: Volume level of air quality pollution.	978.80	734.13
KPI II: Volume level of CO2 pollution.	817.22	684.73
KPI III: Time to obtain information from the air.	246.60	6.55

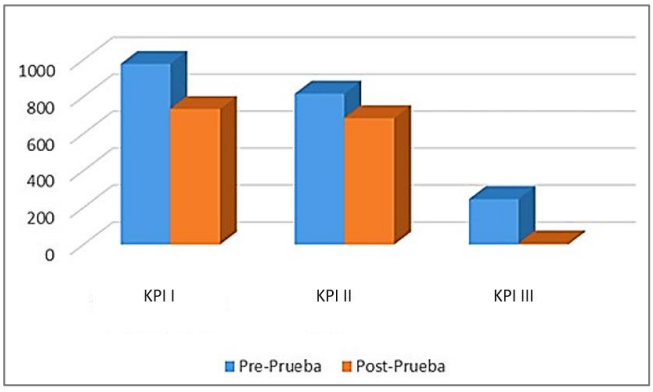


Fig. 8. Comparison of Pre- and Post-test KPI's.

A. Results of the First Indicator (KPI - I)

The summary report of the first post-implementation indicator of the system is shown in Fig. 9.

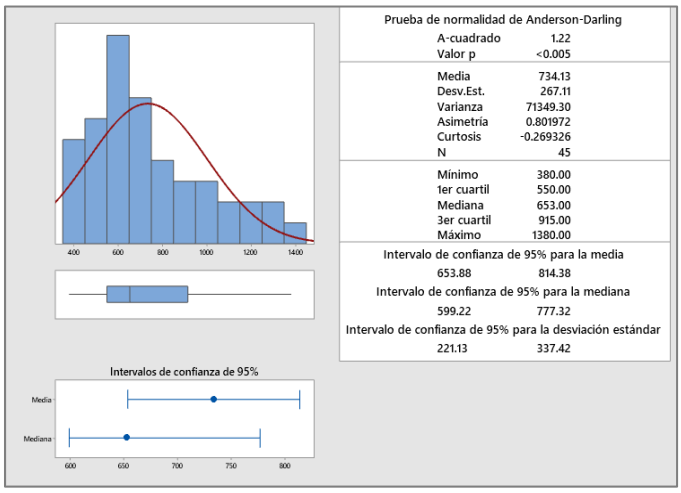


Fig. 9. KPI1 Post-Test Air Quality Pollution Volume Level.

The "average" difference of individual observations of air quality pollution volume level from the mean is 267.11 ppm.

About 95% of the air quality pollution volume level are within 2 standard deviations of the mean, i.e., between 653.88 and 814.38 ppm. The 1st Quartile (Q1) = 550 ppm, tells us that 25% of the volume level of air quality pollution is less than or equal to this value. The 3rd Quartile (Q3) = 915 ppm, tells us that 75% of the volume level of air quality pollution is less than or equal to this value.

Fig. 10 shows the conclusion of the normality test of the KPI-I data obtained from the post-test. The data obtained shows that the p-value is less than 0.05, which confirms that the information analyzed has a non-normal behavior.

B. Results of the Second Indicator (KPI - II)

The summary report of the second Post System Implementation indicator, detailed in Fig. 11, is shown.

The "average" difference of the individual observations of the CO2 pollution volume level relative to the mean is 257.94 ppm.

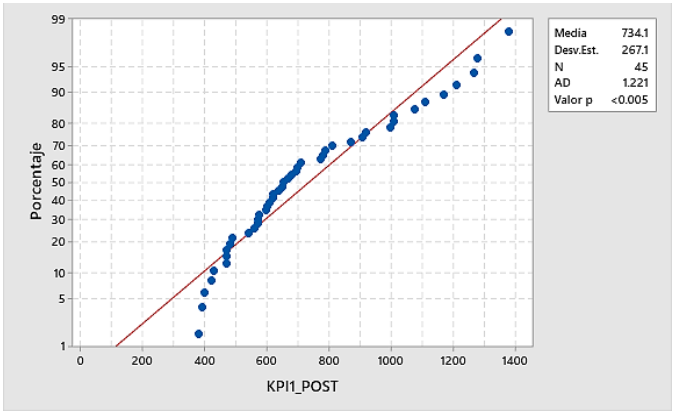


Fig. 10. Normality Test KPI1 Post-Test.

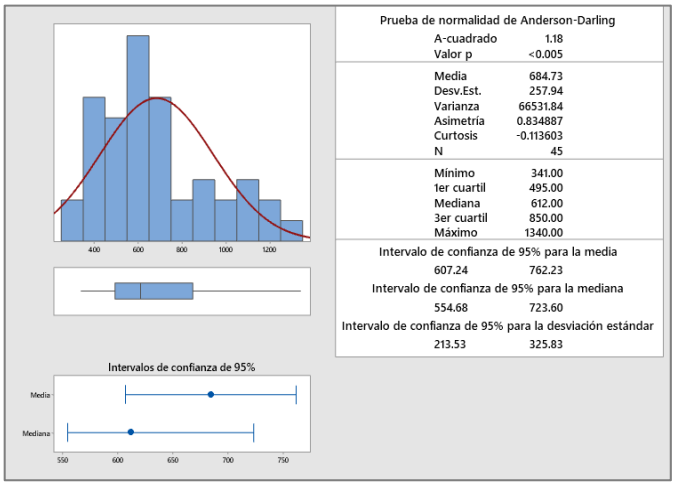


Fig. 11. KPI2 Post-Test CO2 Contamination Volume Level.

About 95% of the CO2 pollution volume level are within 2 standard deviations of the mean, i.e., between 607.24 and 762.23 ppm. The 1st Quartile (Q1) = 495 ppm, tells us that 25% of the volume level of CO2 pollution is less than or equal to this value. The 3rd Quartile (Q3) = 850 ppm, tells us that 75% of the volume level of CO2 pollution is less than or equal to this value.

Fig. 12 shows the conclusion of the normality test of the KPI-II data obtained from the post-test. The data obtained shows that the p-value is less than 0.05, which confirms that the information analyzed has a non-normal behavior.

C. Results of the Third Indicator (KPI - III)

The summary report of the third Post System Implementation indicator, detailed in Fig. 13, is shown.

The "average" difference of the individual observations of the time to obtain air data relative to the mean is 1.25 seconds.

About 95% of the time to obtain information from the air are within 2 standard deviations of the mean, i.e., between 6.17 and 6.93 seconds. The 1st Quartile (Q1) = 5.5 seconds, tells us that 25% of the time to obtain information from the air is less than or equal to this value. The 3rd Quartile (Q3) = 8 seconds, tells us that 75% of the time to obtain information from the air is less than or equal to this value.

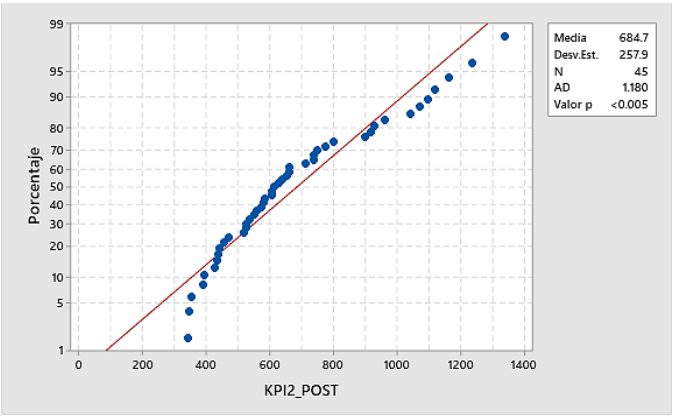


Fig. 12. KP2 Post-test Normality Test.

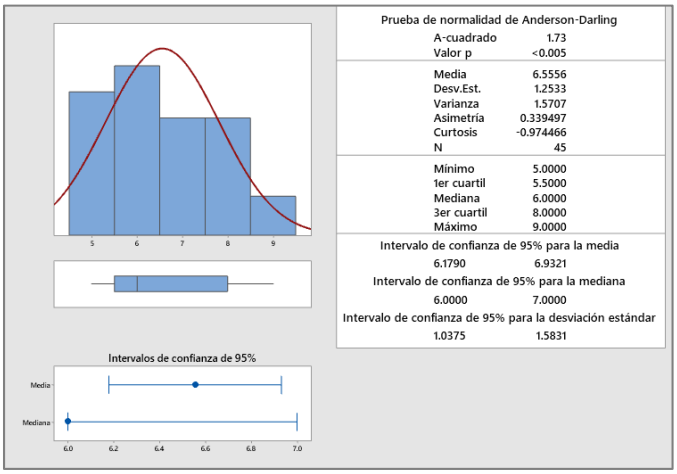


Fig. 13. KPI3 Post-Test Time to Obtain Information from the Air.

Fig. 14 shows the conclusion of the normality test of the KPI-III data obtained from the post-test. The data obtained shows that the p-value is less than 0.05, which confirms that the information analyzed has a non-normal behavior.

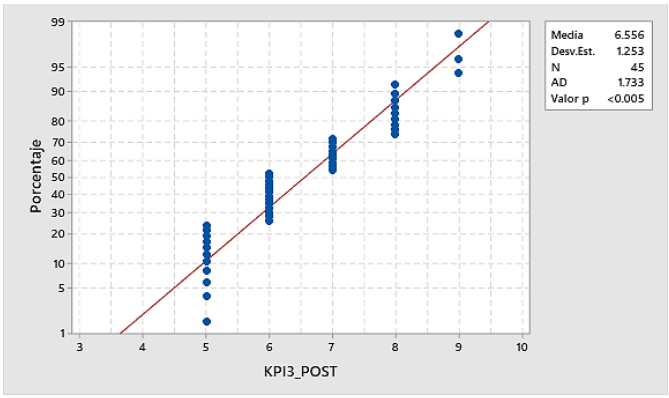


Fig. 14. KP3 Post-Test Normality Test.

V. Discussion

The positive impact of the development of the control system solution on the 3 indicators (Fig. 9, Fig. 11, Fig. 13), carried out on the sample, was verified.

According to Table IV and Fig. 8, it was identified that the use of the control system reduces the volume level of air quality pollution (Post-Test) compared to the sample to which it was not applied (Pre-Test). In addition, the use of the control system reduces the volume level of CO2 contamination (Post-Test) compared to the sample to which it was not applied (Pre-Test). Finally, it was identified that the use of the control system reduces the time to obtain information from the air (Post-Test) compared to the sample to which it was not applied (Pre-Test).

In the present investigation, a confidence level of 95% was established, so a margin of error or significance level of 5% was considered. Considering that the value of $p = 0.000 < \alpha = 0.05$, the resulting data in Fig. 10, Fig. 12, Fig. 14 provide sufficient evidence to reject the null hypothesis, the test being significant.

VI. Conclusion

Finally, it was possible to conclude that the use of the control system evidenced a significant improvement in the air quality of the closed establishment, this is because the polluting gases of the air decreased from 1260 ppm to 670 ppm achieving to be within the standards recommended by WHO; this information was obtained thanks to the sensors installed, in addition to their correct operation in the activation of the ventilation systems of the place, thus fulfilling the proposed objective.

It was possible to verify that with the application developed, the air status collection time was considerably reduced; the time decreased from 350 seconds to 5 seconds since this is constantly shown in the application managed by the manager.

When obtaining the data after the implementation of the system it was possible to verify that the amount of CO2 present in the air decreased, having a CO2 presence of 1150 ppm before the implementation of the system in the air, however, after the use of the system it decreased to 650 ppm, being among the positive indexes.

It was verified that the use of free hardware was able to work correctly, having previously performed the corresponding calibrations and tests of the systems involved in the project, such as sensors and the ESP8266 module.

The scientific contribution provided by this research article is fundamental for the development of future related works serving as a basis, in order to improve the quality of life of people and contribute to the awareness of the damage generated to the environment by the hand of man himself. It is proposed to make improvements in the topic developed through the implementation of new technologies such as artificial intelligence and analysis of data, also make improvements in the hardware and software model as in the case of the application, making it compatible with other operating systems, because betting on developing a hybrid application helps to optimize time, making it possible to generate an application for both Android and IOS.

References

[1] A. Ramos-Romero, B. Garcia-Yataco, and L. Andrade-Arenas, “Mobile Application Design with IoT for Environmental Pollution Awareness,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 1, p. 2021, 2021, doi: 10.14569/IJACSA.2021.0120165.

[2] F. Perera, “Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: Solutions exist,” Int. J. Environ. Res. Public Health, vol. 15, no. 1, 2018, doi: 10.3390/ijerph15010016.

[3] United States Environmental Protection Agency, “Introduction to Indoor Air Quality,” US EPA, 2021. https://www.epa.gov/indoor-air-quality-iaq/introduction-indoor-air-quality.

[4] World Health Organization, “New WHO Global Air Quality Guidelines aim to save millions of lives from air pollution,” WHO, 2021. https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution

[5] Dirección General de Industria; Energía y Minas de la Comunidad de Madrid, “Guía de Calidad del Aire Interior,” Fenercom, p. 186, 2016, Available: https://www.diba.cat/documents/467843/172263104/Guia_qualitat_aire.pdf/eeba42ef-8af3-40e4-b4b3-2f399ed91f31.

[6] L. Barrie and G. Braathen, “Wmo Greenhouse Gas Bulletin: The State of Greenhouse Gases in the Atmosphere Based on Global Observations through 2020,” World Meteorol. Organ., 2020.

[7] K. Lee and M. Greenstone, “Air Quality Life Index | Annual Update,” AQLI, pp. 16–17, 2021. Available: https://aqli.epic.uchicago.edu/pollution-facts/.

[8] A. Ochoa Duarte, L. D. Cangrejo Aljure, and T. Delgado, “Alternativa Open Source en la implementación de un sistema IoT para la medición de la calidad del aire.,” Rev. Cuba. Ciencias Informáticas, vol. 12, no. 1, pp. 189–204, 2018.

[9] R. R. Urquijo and M. J. Marinelli, “Sistema de monitoreo de una cámara de germinación hidropónica con IoT basado en Raspberry Pi,” AGRANDA, Simp. Argentino Gd. Datos, vol. 4, pp. 64–73, 2017.

[10] M. G. Retuerto, D. Y. Espinoza, and L. Andrade-Arenas, “System Dynamics Modeling for Solid Waste Management in Lima Peru,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, p. 2021, 2021, doi: 10.14569/IJACSA.2021.0120762.

[11] World Health Organization, “Ambient (outdoor) air pollution,” WHO, 2021. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[12] R. Perez (WHO-E) and E. Appoh (Ghana-E), “WHO global air quality guidelines,” Germany, 2021. Available: https://cdn.who.int/media/docs/default-source/air-quality-and-health/who-global-aqgs.-afro-presentation-2-nov-2021_final.pdf?sfvrsn=7d2f3da7_5.

[13] World Health Organization, “Household air pollution and health,” WHO, 2021. https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health.

[14] R. UlAmin, M. Akram, N. Ullah, M. Ashraf, and A. Sattar, “IoT Enabled Air Quality Monitoring for Health-Aware Commuting Recommendation in Smart Cities,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 6, 2020, doi: 10.14569/IJACSA.2020.0110637.

[15] J. L. Diaz-Resendiz, A. E. Guerrero-Sanchez, M. Toledano-Ayala, and E. A. Rivas-Araiza, “IoT Based ambient monitoring system for intelligent buildings,” IEEE ICA-ACCA 2018 - IEEE Int. Conf. Autom. Congr. Chil. Assoc. Autom. Control Towar. an Ind. 4.0 - Proc., pp. 1–6, 2019, doi: 10.1109/ICA-ACCA.2018.8609862.

[16] A. Delgado et al., “Comparative Analysis of the Impact on Air Quality Due to the Operation of La Oroya Metallurgical Complex using the Grey Clustering Method,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 2, pp. 450–454, 2021, doi: 10.14569/IJACSA.2021.0120257.

[17] V. Hable-Khandekar and P. Srinath, “Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring,” 2017 Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2017, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2017.8463746.

[18] Y. C. Wang and G. W. Chen, “Efficient Data Gathering and Estimation for Metropolitan Air Quality Monitoring by Using Vehicular Sensor Networks,” IEEE Trans. Veh. Technol., vol. 66, no. 8, pp. 7234–7248, 2017, doi: 10.1109/TVT.2017.2655084.

[19] S. Mansour, N. Nasser, L. Karim, and A. Ali, “Wireless sensor network-based air quality monitoring system,” 2014 Int. Conf. Comput. Netw. Commun. ICNC 2014, pp. 545–550, 2014, doi: 10.1109/ICCNC.2014.6785394.

[20] J. Dutta, F. Gazi, S. Roy, and C. Chowdhury, “AirSense: Opportunistic crowd-sensing based air quality monitoring system for smart city,” Proc. IEEE Sensors, pp. 5–7, 2017, doi: 10.1109/ICSENS.2016.7808730.

[21] A. Tapashetti, D. Vegiraju, and T. Ogunfunmi, “IoT-Enabled Air Quality Monitoring Device,” IEEE 2016 Glob. Humanit. Technol. Conf., pp. 0–3, 2016.

[22] A. A. Alkandari and S. Moein, “Implementation of Monitoring System for Air Quality using Raspberry PI: Experimental Study,” Indones. J. Electr. Eng. Comput. Sci., vol. 10, no. 1, p. 43, Apr. 2018, doi: 10.11591/ijeecs.v10.i1.pp43-49.

[23] H. F. Hawari, A. A. Zainal, and M. R. Ahmad, “Development of real time internet of things (IoT) based air quality monitoring system,” Indones. J. Electr. Eng. Comput. Sci., vol. 13, no. 3, p. 1039, Mar. 2019, doi: 10.11591/ijeecs.v13.i3.pp1039-1047.

[24] S. Pandey, S. H. Saeed, and N. R. Kidwai, "Simulation and optimization of genetic algorithm-artificial neural network based air quality estimator," Indones. J. Electr. Eng. Comput. Sci., vol. 19, no. 2, p. 775, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp775-783.

[25] J. Patricia, Z. Gamboa, C. Alexandra, and L. Arreaga, "Evolución de las Metodologías y Modelos utilizados en el Desarrollo de Software. Evolution of the Methodologies and Models used in Software Development," INNOVA Res. J., vol. 3, no. 10, pp. 20–33, 2018.

Using HBase to Implement Speed Layer in Time Series Data Storage Systems

Milko Marinov
Department of Computer Systems and Technologies
University of Ruse, Ruse, 7017, Bulgaria

***Abstract*—In recent years, modern systems have become increasingly integrated, and the challenges are focused on delivering real-time analytics based on big data. Thus, using standard software tools to extract information from such datasets is not always possible. The Lambda Architecture proposed by Marz is an architectural solution that can manage the processing of large data volumes by combining real-time and data batch processing techniques. Choosing a suitable database management system for storing large volumes of time series data is not a trivial issue as various aspects such as low latency, high performance and the possibility of horizontal scalability must be taken into account. The new NoSQL approaches use for this purpose non-relational databases with significant advantages in terms of flexibility and performance in comparison with the traditional relational databases. With reference to this, the purpose of this paper is to analyse the general characteristics of time series data and the main activities performed by the Speed layer in a system based on the Lambda Architecture. Based on this, the use of a column-oriented NoSQL DBMS as a system for storing time series data is justified. The paper also addresses the challenges of using HBase as a system for storing and analysing time series data. These questions are related to the design of an appropriate database schema, the need to achieve balance between ease of access to the data and performance as well as considering the factors that affect the overload of individual nodes in the system.**

Keywords*—*Lambda architecture; speed layer; time series data; data storage system

I. Introduction

The accelerated development of technologies applied to big data has caused significant changes in the subject areas of storage, retrieval, and processing of data. Nowadays, the problems related to the big data are connected not only to the volume of data. Much of the data are acquired in real time and is most valuable if its interpretation takes place as it arrives [1,2,3]. Synthesizing, processing, and transforming this big data to valuable information is one of the great challenges of the technological world today.

In big data systems, an important property of data related to its processing is immutability. In such systems, to prevent data loss and data corruption, data are processed in a way that records can never be modified or deleted. By its nature, immutable data are simpler than mutable data [4]. This organization allows the system only to create and read records (CreateRead) as opposed to the additional capability of updating and deleting records (CRUpdateDelete) as implemented in relational databases. Thus, the write operations only add new data units [5]. This approach makes data processing highly scalable. The data system itself becomes a kind of a logging system, which adds a timestamp and a unique identifier to the data record, which is then kept in the data store.

Different architecture models are used when building the big data ecosystem. The Lambda Architecture, proposed by Nathan Marz [6], is a solution which combines real-time data processing techniques with batch processing techniques. The Lambda Architecture is a big data management software paradigm that supports data processing by balancing the performance, latency, and fault-tolerance of the system that is based on this architecture [7,8]. There is no single integrated tool that provides a comprehensive solution with reference to better accuracy, low latency, and high performance. Therefore, it is necessary to apply the idea of using a set of tools and techniques to build a comprehensive big data management system. With reference to this, Lambda Architecture defines several layers that correspond to a set of tools and techniques for building a big data processing system, i.e., a speed layer, a serving layer, and a batch layer [9]. The increasing need for new and improved storage and retrieval mechanisms resulted in the development and use of NoSQL database management systems such as HBase, MongoDB, Cassandra, CouchDB, Hypertable and big data platforms such as Hadoop and Spark [10,11,12]. Lambda Architecture defines a logical, well-motivated approach in linking these technologies together to build a system that meets user requirements. Each software tool offers its own trade-offs, but when these tools are used together, scalable systems with low latency, high fault-tolerance, and minimal complexity can be realized [13].

The main objective of the current research is to justify the use of a column-oriented NoSQL DBMS as a system for storing time series data. This suggestion is based on the general characteristics of time series data and the main activities performed by the Speed layer in a Lambda Architecture based system. This paper focuses on the challenges of using HBase as a system for storing and analysing time series data, related to designing an appropriate database schema, the need to strike a balance between ease of data access and performance, and consideration of factors that affect the overload of individual nodes in the system.

The remainder of this article includes the following: Section 2 surveys some related studies. Section 3 presents an overview of the Lambda Architecture with an emphasis on the Speed layer. Section 4 discusses the main characteristics of time series data. Section 5 outlines the key problems that arise

when using HBase as a system for storing time series data and techniques for solving these issues. Section 6 contains the conclusion.

II. Related Work

One of the big challenges to extracting data from processes is handling real-time event data and providing operational support for ongoing processes. The research presented in [7] focuses on a real-time process discovery algorithm implemented by the authors in an integrated platform that is built using the Lambda Architecture principles. The proposed architecture solution makes it possible to scale up to big data processing tasks.

Trajectory prediction problems are classified in the category of big data processing tasks. In [1], a platform for predicting the next position of moving objects based on the Lambda Architecture is discussed. In the presented platform, data analysis is performed both in a batch mode and a real time mode. In the proposed system, the Lambda Architecture is applied to combine predictions made by heavy-weight models trained by using all available data, implemented by the Batch layer, on one hand, and light-weight models trained by using real-time data obtained from small samples, implemented by the Speed layer, on the other hand.

Maeda & Gaur propose in [14] a Lambda Architecture of a failure mode identification system for industrial assets that achieves low initial implementation costs by providing reasonable accuracy in object classification. The architecture consists of a data acquisition node, such as a Raspberry Pi, in which lightweight computations are performed. This node processes high-speed vibration data in real-time to extract important characteristics about objects and uses a deep learning engine that is trained in a cloud platform.

The nature of heterogeneous IoT devices introduces the challenge of collecting and processing the large data sets for their analysis in detecting cyber-attacks in near real-time. However, the traditional Intrusion Detection System cannot cope with such a problem due to scalability limitations and insufficient storage and processing capabilities. To address these challenges, Alghamdi & Bellaiche present in [15] a model of Intrusion Detection System based on the Lambda Architecture. The proposed solution enables the detection of suspicious activities in real time and allows them to be classified by analyzing historical data in the Batch layer. Suthakar et al. describe in detail in [9] a study of an Optimized Lambda Architecture using the Apache Spark ecosystem, which involves the modelling of an efficient way to transparently connect batch processing and real-time processing.

Data storages, whose architecture is based on the relational data model, are not able to meet the current needs in terms of data storage requirements as well as data read and write speed requirements. Therefore, research related to real-time data management systems is based on distributed storage organized in a cluster, for example built on the Hadoop ecosystem. The study in [4] proposes the use of HBase DBMS, whose storage structure is based on the column-oriented data model. The described system provides real-time monitoring of sensor data and satisfies data storage and processing requirements. In [16], the researchers discuss the use of the Hadoop ecosystem in finance.

Bao & Cao analyse in [17] the challenges to storing and retrieving social network data. In this study, they present a query optimization scheme based on HBase DBMS. The table structure is designed according to the characteristics of HBase, such as the high efficiency of row keys and storage which is based on the column-oriented data model. A coprocessor is used to design a secondary index, which transforms some queries to the attributes into row key queries of the index tables so that this can support flexible queries to social network data with high scalability and low latency. In [18,19], a similar solution is proposed regarding the indexing mechanism in HBase for sensor data processing. A secondary memory index mechanism is used. The retrieval speed of the indexed data is significantly improved because the indexing is stored and maintained in the memory.

In order to increase the performance of the storage process of data retrieved in the real time, the process must be distributed. It can be realized by taking advantage of the MapReduce technology. HBase offers a data loading tool that can process data stored in TSV or CSV formats which is called ImportTSV [20]. This tool is based on the MapReduce model. Azqueta-Alzúaz et al. in [10] identify and quantify the problems associated with loading massive big data. They propose a tool for parallel massive data loading using HBase. This solution overcomes the defined problems of data loading in HBase.

III. Lambda Architecture Overview

The CAP theorem applies to trade-offs in distributed systems [6,11]. It states that in a distributed data storage system only two of the characteristics availability, consistency, and partition tolerance can be guaranteed. The meaning of partition tolerance is that the system characteristics are maintained even in case of network failures. This requirement is supported in modern systems. Therefore, when designing a data management system, a compromise has to be made between consistency and availability. The Lambda Architecture is designed to address the trade-offs that must be made in a distributed data storage and processing system. The main idea is to create two input data streams to be processed separately and to combine later the obtained results. The components of the Lambda Architecture are the Batch layer, the Serving layer, and the Speed layer (Fig. 1).

The components through which the Batch layer and the Serving layer are implemented involve the execution of computational functions on each piece of data, i.e., on the data as a whole [6]. These layers satisfy all characteristics that are required for a data processing system except one: low latency updates. The only task of the Speed layer is to satisfy the latter requirement. Executing functions on an entire data set, which could be measured in petabytes, is an operation that requires considerable computational resources. The Speed layer should use a completely different approach from the one used by the Batch and Serving layers to reduce latency of updates as much as possible. Whenever data changes, the Speed layer recalculates only those results that depend on the changed data,

i.e., incremental computation is performed. The main functionalities that must be implemented by the Speed layer are storing real-time views and processing the incoming data stream to update these views. The Speed layer is more complex than the Batch layer and more errors may occur when storing and processing the data. The Speed layer is only responsible for the data that will be included in the Batch views, which are part of the Serving layer. The amount of this data is significantly smaller than the amount of the main dataset. This allows more flexibility in the design of the Speed layer. Real time views that are obtained as a result of the Speed layer are temporary, i.e., they are not stored permanently. Once the data are accepted and used in the batch views, it can be discarded from the Speed layer.

The choice of Lambda Architecture when building a distributed system may have the following disadvantages:

- The different layers in this architecture make it too complex in terms of synchronization between layers. It is possible that the cost of synchronization between the batch and speed layer will increase. More computational resources, time and efforts are required to run both the Batch and Speed layers.
- The implementation of the Lambda Architecture requires the use of a large number of technologies, which makes it difficult to find specialists who know the whole set of tools.
- Under certain conditions, this architecture could contain a large surplus of tools that need to be configured for each scenario.

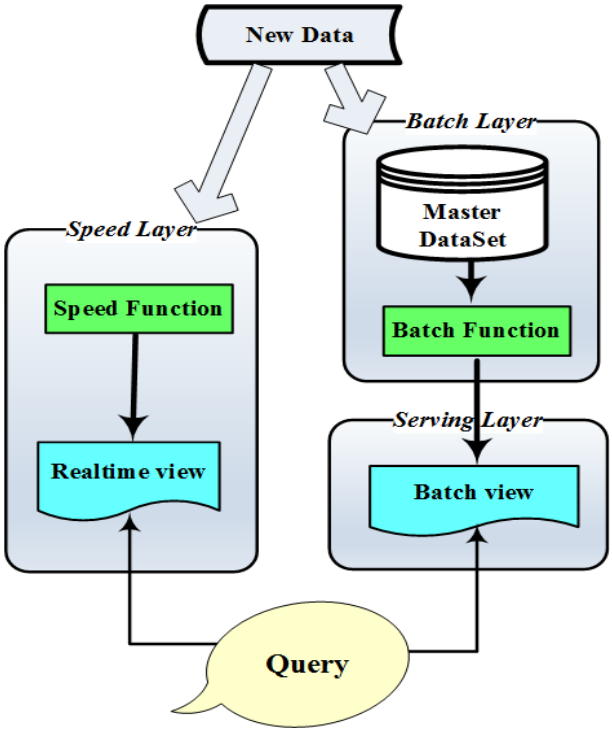


Fig. 1. Lambda Architecture Main Components.

IV. Characteristics of Time Series Data

Time series data are a set of data points, and each value is connected to a timestamp. Formally, a time series can be defined as a set of pairs, each consisting of a timestamp and a value [2,20]. Generally speaking, time series datasets are sequences of data records ordered according to the time of their occurrence. Time series data is characterized by chronological consistency, large volume and high degree of competitiveness. The time series datasets are used in situations in which, once measurements have been taken, they are not revised or updated. However, they are rather used when the set of measurements is accumulated by adding new data for each parameter that is measured at each new time point. Time series data entries are rarely changed, and time series data are extracted by reading a continuous sequence of samples, obtained after summarizing or aggregating the samples, in the order in which they are received. These time series data characteristics limit the requirements for the technology used to store this type of data [23].

Although the idea of collecting and analyzing time series data is not new, the combination of some factors such as the volume of current datasets, the speed of data accumulation, and the variety of new data sources turn the task of building scalable time series databases into an enormous challenge. Time series data require different approaches and different tools [24].

It is difficult to store data whose nature is unpredictable. This makes it necessary to store and process time series data in a database. Especially when it comes to large volumes of time series data, the requirements for its efficient storage become increasingly important. A time series database is a way of storing multiple time series so that queries to retrieve data from one or more time series for a specific period can be executed efficiently. Time series databases allow users to predict the behavior of an object by analyzing its past states. The queries made to the time series data can be implemented as large, sequential scans, which are very efficient if the data are stored appropriately in a time series database. And if the data volume is very large, a non-relational time series database based on a suitable NoSQL data model is usually needed to provide sufficient scalability. In addition to considering the characteristics, the nature of time-series data, as well as the requirements for high storage reliability and horizontal scalability, require the use of a NoSQL distributed time-series database as the most suitable for storing and processing all these large volumes of data. The new NoSQL-based approaches use non-relational databases for this purpose with significant advantages in terms of flexibility and performance over traditional relational databases.

V. Storing and Processing Time Series Data in HBase

HBase is a distributed DBMS built on HDFS. One of the significant advantages of HBase is the ability to combine real-time queries with batch MapReduce jobs in the Hadoop ecosystem, using HDFS as a shared storage platform [21,22]. All rows in HBase are sorted lexicographically by row key. In the column-oriented model, the data are organized at the logical level into tables, rows, and columns. A table in HBase

is multi-dimensional and can be queried using the primary key. The key structure is presented in Fig. 2. HBase columns can have multiple versions of the same row key. A typical HBase cluster has one active master node, one or more backup masters, and regional servers (Fig. 3). The HBase Master node assigns the corresponding regions to Region Servers. The first one is the ROOT region, which contains all the META regions which must be assigned. It also monitors the state of Region Servers and if it detects a failure in any Region Server, it restores it using the replicated data. In addition, the HBase Master is responsible for table maintenance. The tasks which HBase Master performs are related to adding or deleting tables and making changes to the table structure. The Region Server handles read and write client requests. It interacts with the HBase Master to obtain a list of regions to serve and informs the master node of its availability.

When HBase is used as a system for storing time series data, problems arise, and they are related to overloading one of the Range Servers and scattering of data. Since there is a wide variety of time series data, it is necessary to take into account the specific features of every type of data when it is stored in HBase.

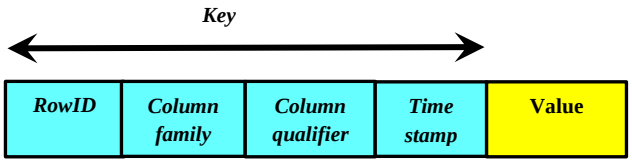


Fig. 2. Key Structure in Column-oriented Data.

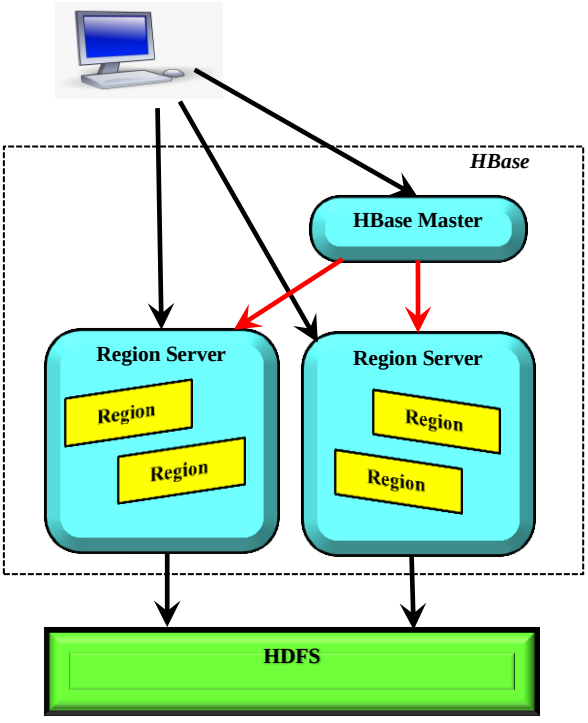


Fig. 3. HBase General Architecture.

The most logical key that can be used to store time series data is the timestamp. This guarantees the uniqueness of the key for every measurement at a specific point in time. With this approach to data storage, the data for each timestamp value can be accessed by performing a single read operation. It is also possible to easily perform a scan of a range of key values. A fast scan is guaranteed because all rows stored in HBase are sorted by key. However, the following two problems arise when this approach is applied and either of them can reduce the system performance:

- First, such an organization of the data may cause overloading of some of the Region servers. This is because at the time of writing, all data are concentrated in the regions that serve the corresponding key values, while in the other regions no data are being written. Similarly, when performing a read query targeting the most recent data, a small number of regions will be accessed. This will reduce the effect of being able to access a larger number of regions in parallel.
- Second, relatively few columns are stored in each row, and this can be very inefficient because little data are read at once and there are too many Bloom filter values that will be used in the search process.

When all new rows are written sequentially in HBase, they are all placed on the same server because they are sorted, and this requires them to be close together. HBase has a built-in automatic sharding mechanism. The new regions (areas of the hard disk where the data are written) which result from the sharding operation will be used later. In this way, they will balance the overload. In practice, the overload will not be noticeable at low write speeds as the RegionServer will be doing perfectly fine. This, however, will not lead to the efficient use of the entire HBase cluster because only one server will be used.

To avoid key concentration in sequential writes, it is necessary to do one of the following when developing the logical organization of the data:

- Indexing must be avoided if possible. In the case of time series data, the timestamp should not be used as the only key, some other data element should be added as well. In other words, a composite key should be used to uniquely identify the row and the moment when the corresponding value was received, or the event occurred.
- Storing write operations randomly. The main problem with time series data is that the sorted timestamps are the essential information, and they must be present in the data in one or another way. Therefore, using random writes in time series data will make the writes faster but the reads slower because the data will have to be collected from many locations. In some cases, even pseudo randomization can help to reduce the load of some servers.

- Adding a fragment identifier at the beginning of the key. In this way, the load can be distributed among the set of Region servers. When the data are being read, it must be read by each server. Then the fragment identifier must be added to the timestamp, and finally the query results must be combined in the memory.

If the write of a time series data set has a rather small row (with a small number of columns), this will lead to a problem associated with cache-hit and large Bloom filters when data are searched for. Cache-hit is a condition in which the data that is requested for processing is in the cache memory. In terms of HBase, this can be explained as follows. Hadoop reads blocks from HDFS, which typically range in size from 64 MB to 256 MB. If the data to be read and written is much smaller than this size, this will lead to inefficient cluster operation and hence cache memory problems. Bloom filters are used to answer the question whether, based on the key, it is possible to locate the data in the corresponding region. The answer is not definite, but it should be understood as maybe yes, which requires reading the region and searching. If the keys identify rows containing little information (called thin rows), this will cause the use of too many Bloom filter keys, which will take up disk space and reduce the efficiency of their use.

One technique for increasing the speed at which data can be retrieved from a time series database is by storing a large number of values in each row. In DBMSs that support a column-oriented data model, and HBase is exactly such a system, the number of columns is almost unlimited. This feature can be used to store numerous values within a single row. This allows data points to be accessed at a higher speed. The speed at which data can be scanned depends on the number of rows scanned, the total number of values retrieved, and the volume of data retrieved. If the number of rows is reduced, the fraction of data loss in retrieval will be significantly reduced, resulting in an increase in retrieval speed. For example, if the row key contains <time_series_ID> and <Start_time_of_Tme_Window>, and the column names correspond to the offset from the start of the time window when the value of the corresponding data element will be written, then the result will be a table with many columns. This means that the data retrieval from a particular time series for a particular time period would involve mainly sequential read operations and would therefore be much faster in comparison to a situation in which the rows were scattered.

Such an organization leads to a reduction in the number of rows in the table. In addition, rows that contain data from the same time series are close to each other when the data are stored. To take advantage of the benefits of this structure with reference to its performance, the number of samples in each time window must be sufficiently large. This will cause a significant reduction in the number of rows that must be retrieved.

This technique is similar to the default table structure used by OpenTSDB [5,20]. OpenTSDB is an open-source distributed time series database designed to control clusters of commodity servers with a high level of granularity. The interaction between OpenTSDB and HBase is presented in Fig. 4. An OpenTSDB consists of interacting components for loading and accessing time series data. These include data collectors, Time Series Daemons (TSDs), and various user interface management related functions. Each TSD is independent. There is no master, no shared state, and as many TSDs as required can be run so that the system can handle the workload. Each TSD uses HBase to store and retrieve time series data. On the servers where measurements are being taken, there is a collection process that sends data to the TSD. The TSDs are responsible for finding time series to which data will be added and each data point will be inserted as it is received in the data storage layer. OpenTSDB uses HDFS as a file system for storing large data sets. A simplified web-based user interface is supported, and users query various metrics in real time through it.

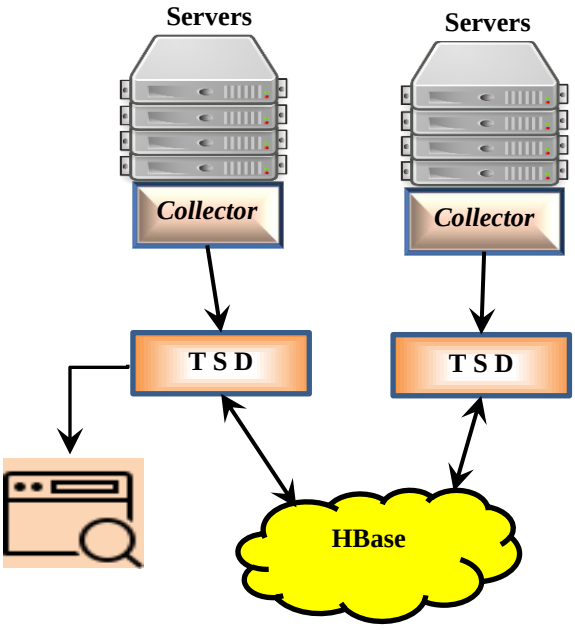


Fig. 4. OpenTSDB Components Interaction.

VI. Conclusion

The Lambda Architecture allows users to optimize the cost of processing large volumes of data by dividing the storage and processing of input data into two streams - data that needs to be processed in real time and data on which batch processing will be performed. The Lambda Architecture provides a consistent approach to building a big data system that can perform real-time data storage and processing in a low-latency, high-throughput, and fault-tolerant manner. The Speed layer implementation needs to consider the characteristics of time series data as well as the requirements of high storage reliability and horizontal scalability. This requires the use of a NoSQL distributed time series database based on a column-oriented data model.

When using HBase as a time series data storage system, it is necessary to properly design a row key that is based on the timestamp in order to overcome the problems associated with overloading one of the Range Servers and the data scatter problem. The effectiveness of the speed layer can be

significantly increased when HBase is integrated with OpenTSDB. All OpenTSDB data points are stored in one “big” table, which is called tsdb by default. All values are stored in a single column family. This is done to take advantage of the key ordering in HBase and the distribution of regions over individual RegionServers.

The author's further efforts will be focused on expanding research on the application of other software architectures used to build time series data storage systems and incorporate these technologies in the Distributed Databases course of the Computer Systems and Technologies master degree at the University of Ruse.

REFERENCES

[1] E. Psomakelis, K. Tserpes, D. Zissis, D. Anagnostopoulos and T. Varvarigou, “Context agnostic trajectory prediction based on λ-architecture,” Future Generation Computer Systems, vol. 110, pp. 531–539, 2020.

[2] A. Noury and M. Amini, “An access and inference control model for time series databases,” Future Generation Computer Systems, vol. 92, pp. 93–108, 2019.

[3] A. Pandya, O. Odunsi, C. Liu, A. Cuzzocrea and J. Wang, "Adaptive and efficient streaming time series forecasting with Lambda architecture and Spark," in 2020 IEEE International Conference on Big Data (Big Data), pp. 5182-5190, 2020.

[4] J. Yang, X. Chi, M. Zhu and W. Wang, “Research and Design of Sensor Data Management System Based on Distributed Storage,” in 2020 Int. Conf. on Computer Science and Management Technology (ICCSMT), pp. 128-132, 2020.

[5] A. Nielsen, Practical Time Series Analysis, O'Reilly Media, Inc., CA, 2019.

[6] N. Marz and J. Warren, Big Data: principles and best practices of scalable real-time systems, Manning Publications, 2015.

[7] A. Batyuk and V. Voityshyn, “Streaming Process Discovery for Lambda Architecture-Based Process Monitoring Platform,” IEEE 13th Int. Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), pp. 298-301, 2018.

[8] M. Kiran, P. Murphy, I. Monga, J. Dugan and S. S. Baveja, "Lambda architecture for cost-effective batch and speed big data processing," 2015 IEEE International Conference on Big Data (Big Data), pp. 2785-2792, 2015.

[9] U. Suthakar, L. Magnoni, D. R. Smith and A. Khan, “Optimised Lambda Architecture for Monitoring Scientific Infrastructure,” in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 6, pp. 1395-1408, 2021.

[10] A. Azqueta-Alzúaz, M. Patiño-Martinez, I. Brondino and R. Jimenez-Peris, “Massive Data Load on Distributed Database Systems over HBase,” 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 776-779, 2017.

[11] D. McCreary and A. Kelly, Making sense of NoSQL, Manning Publications, 2014.

[12] B. Jose and S. Abraham, “Performance analysis of NoSQL and relational databases with MongoDB and MySQL”, Materials Today: Proceedings, vol. 24(3), pp. 2036-2043, 2020.

[13] F. Cerezo, C. E. Cuesta, J. C. Moreno-Herranz and B. Vela, "Deconstructing the Lambda Architecture: An Experience Report," 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), pp. 196-201, 2019.

[14] D. Maeda and S. Gaur, “Lambda architecture for robust condition based maintenance with simulated failure modes,” IEEE/ACM Symposium on Edge Computing (SEC), pp. 152-154, 2020.

[15] R. Alghamdi and M. Bellaiche, “A Deep Intrusion Detection System in Lambda Architecture Based on Edge Cloud Computing for IoT,” 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 561-566, 2021.

[16] F. de Moura Rezende dos Santos and M. Holanda, “Performance Analysis of Financial Institution Operations in a NoSQL Columnar Database,” 15th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-6, 2020.

[17] C. Bao and M. Cao, “Query Optimization of Massive Social Network Data Based on HBase,” IEEE 4th International Conference on Big Data Analytics (ICBDA), pp. 94-97, 2019.

[18] F. Ye, S. Zhu, Y. Lou, Z. Liu, Y. Chen and Q. Huang, "Research on Index Mechanism of HBase Based on Coprocessor for Sensor Data," IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), pp. 598-603, 2019.

[19] P. Zhengjun and Z. Lianfen, "Application and research of massive big data storage system based on HBase," IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 219-223, 2018.

[20] T. Dunning and E. Friedman, Time Series Databases: New Ways to Store and Access Data, O’Reilly Media, Inc., CA, 2015.

[21] P. Wang, F. Xu, M. Ma and L. Duan, "Efficient Spatial Big Data Storage and Query in HBase," 2019 IEEE International Conference on Smart Cloud (SmartCloud), pp. 149-155, 2019.

[22] H. Ochiai, H. Ikegami, Y. Teranishi and H. Esaki, "Facility Information Management on HBase: Large-Scale Storage for Time-Series Data," 2014 IEEE 38th International Computer Software and Applications Conference Workshops, pp. 306-311, 2014.

[23] G. Liu, W. Zhu, C. Saunders, F. Gao and Y. Yub, “Real-Time Complex Event Processing and Analytics for Smart Grid”, Procedia Computer Science, vol. 61, pp. 113-119, 2015.

[24] K. Mishra, S. Basu, U. Maulik, “Graft: A graph based time series data mining framework,” Engineering Applications of Artificial Intelligence, vol. 110, 2022.

Machine Learning Model for Prediction and Visualization of HIV Index Testing in Northern Tanzania

Happyness Chikusi, Dr Judith Leo, Dr Shubi Kaijage
School of Computational and Communication Science and Technology
Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania

***Abstract*—Human Immunodeficiency Virus Acquired Immunodeficiency Syndrome (HIV AIDS) in Tanzania is still a threatening disease in society. There have been various strategies to increase the number of people to know their HIV status. Among these strategies, HIV index testing has proven to be the best modality for collecting the number of HIV contacts who might be at risk of contracting HIV from an HIV-positive person. However, the current HIV index testing is manual-based, creating many challenges, including errors, time-consuming, and expensive to operate. Therefore, this paper presents the Machine Learning model results to predict and visualise HIV index testing. The development process followed the Agile Software development methodology. The data was collected from Kilimanjaro, Arusha and Manyara regions in Tanzania. A total of 6346 samples and 11 features were collected. Then, the dataset was divided into training sets of 5075 samples and a testing set of 1270 samples (80/20). The datasets were run into Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN) algorithms. The results of the evaluation, by Mean Absolute Errors (MAE), showed that; RF MAE (1.1261), XGBoost MAE (1.2340), and ANN MAE (1.1268.); whereby the RF appeared to have the best result compared to the other two algorithms. Data visualisation shows that 17.4% of males and 82.6 of females had been notified. In addition, the Kilimanjaro region had more cases of people with HIV status from their partners. Overall, this study improved our understanding of the significance of ML in the prediction and visualisation of HIV index testing. The developed model can assist decision-makers in coming out with a suitable intervention strategy towards ending HIV AIDS in our societies. The study recommends that health centres in other regions use this model to simplify their work.**

Keywords*—Index testing; machine learning; random forest; XGBoost; artificial neural network

I. Introduction

Index testing refers to a case-finding strategy that aims to get the exposed contacts of HIV Positive individuals for HIV-testing services. It is also known as partner notification [1]. This person is known as an indexing client. Healthcare workers and counsellors ask index clients to list all their partners, including sexual partners and or injecting drugs partners and their children. The process is voluntary and confidential. In the process, each partner and the children are contacted and informed on the exposure to HIV and offered voluntary testing. The purpose of index testing is to break the chain of HIV transmission. In addition, health workers provide HIV testing to the people who have been exposed to HIV[2]. If the result is positive, they are linked to the treatment, and if the status is negative, they are given prevention services.

There are various HIV testing modalities such as Voluntary HIV counselling and Testing (VCT), Community VCT home-based, mobile, and outreach testing. However, home-based and mobile outreach is costly. Therefore, index case testing was introduced to increase the number of people to know their status, and it has been a promising strategy towards the maximisation of HIV case detection [3]. The current HIV index client testing system does not have an automated system, and the data are collected manually. Therefore, it is challenging to analyse the data and predict HIV index testing. In addition, it requires expertise for data entry and data analysts to do the work. Hence, resulting in additional cost and time-consuming in obtaining the intended results. Not only that but also human errors are unavoidable.

Other researchers applied machine learning in health care specific to HIV AIDS solving various problems like HIV case findings, HIV predictors, Patient-specific current CD_4 count, and prediction of new HIV Index using internet data, however, the methods used were statistical descriptive, estimated index and chi-square.

Therefore, this paper presents the results of the developed Machine Learning model that can help experts make predictions and produce up-to-date data visualisation that is readable and understandable. In addition, the developed Machine-learning model can predict the number of HIV Index testing using partner notification information to identify people who are at risk to contract HIV AIDS. Hence, help decision-makers to come out with a good intervention strategy towards ending HIV AIDS in our societies.

The paper consists of five parts namely: introduction, literature review, material and methodology, results and discussion, conclusion and recommendation.

II. Literature Review

A. Overview of Literature Survey

HIV is an infectious disease that threatens public health globally. According to World Health Organization (WHO), 38% million people are living without HIV globally. 19% do not know their status[4]. Many people living with HIV are

located in middle and low-income countries, with an estimated 68% in sub-Saharan Africa.

WHO Strategy of 2016 to 2021 addresses human rights and equity, with a radical decline in a new HIV infection and reducing death. The global target is to reduce new infection to less than 500,000 by 2020 and end HIV by 2030 as a public threat. The current target is 90 90 90, meaning that 90% know their status, 90% receives quality treatment and care, and the last 90% retain in extended care [5][6].

In the southern part of Africa, various countries have made substantial progress toward the HIV/AIDS Program target of ensuring that 90% of people living with HIV know their status. HIV testing and counselling is the crucial step towards achieving the Joint United Nations Program on HIV/AIDS (UNAIDS) of 90 90 90. However, the target for 2025 is 95% 95% 95%[7].

B. HIV Trends in Tanzania

The HIV status in Tanzania shows that 1.7million people live with HIV, 77,000 new HIV infections, and 27,000 AIDS-related death[8]. The new strategic plan reviewed by the Ministry of Health for 2018 to 2022 is making Index testing Services and Partner Notification services one of the National Strategy for Identification of the People Living with HIV (PLHIV)[9]. Index case testing will support Tanzania to maximise HIV case detection in achieving the first target of 90 (2017- 2022) and the next of 95 for 2025 for males, adolescents, and children.

C. Machine Learning in Health Care

Machine learning(ML) is the use and development of computer systems that can learn and adapt without following explicit instructions use algorithms to analyse and draw inferences from the pattern in data[10]. Machine learning algorithms depend on domain knowledge of the data to create features that make these algorithms work. ML has been used in various domains with data availability, including computer vision, automatic speech recognition, business analysis, natural language processing, and even health care. However, the process demands lots of time and effort for feature selection, and features must extract relevant information from vast and diverse data to produce the best outcome.

Machine learning techniques accurately provide predictions in various applications, such as drug discovery and disease diagnosis, especially with quality data. Machine learning interest is in cancer diagnosis, diabetes, autism subtyping in health care[11]. Also, ML is used to predict cholera disease [12].

Machine learning in HIV/AIDS had applied as follows: Machine learning to identify HIV predictors for screening [13]. Machine learning in the prediction of patient-specific current CD_4 cell count to determine the progression of human immunodeficiency.[14] Prediction of new HIV infection in China by using internet search [15], predicting default from HIV service in Mozambique [16], Another area is improving HIV case findings [17].

Other related works predict HIV index Testing using different methods are Index and target community testing to optimise HIV case findings among men. The process used descriptive statistics, estimated index cascade, and Chi-Square test.[18] Sustained high HIV case finding through Index testing via services register using Microsoft excel.[19] Another study done was about applying machine learning on HIV/AIDS diagnosis and therapy planning[20].

Therefore, this study aims to use machine-learning techniques to predict HIV index testing and visualisation items of Age, Sex, location, and relationship to strengthen the ability to plan, prioritise, and implement the effective intervention.

III. Materials and Methods

A. Materials

The study area selected was the northern part. The Northern party regions include Tanga, Kilimanjaro, Arusha, and Manyara. Kilimanjaro, Arusha, and Manyara had chosen to represent the party. The dataset used in this study was from different health centres and community sites from Arusha, Kilimanjaro, and Manyara. The client information consists of 6346 samples and 11 features, and index-client data consists of 7226 samples with 13 elements.

Python was the programming language used in this study. The reason that led to this programming language took into consideration its ability to offer a variety set of open-source libraries to support machine learning.

B. Methoderations

1) Knowledge discovered from data science: Knowledge Discovered from Data (KDD) is extracting knowledge from various vast quantities of data. In carrying out this study, we selected this approach due to its application in data mining using different algorithms and clearly defined phases. [21]. The study followed an interactive refine at each step (Table I) explains the stages of KDD.

TABLE I. Summary of KDD

NAME	EXPLANATION
Problem Identification	It is the bedrock of all the stages. It involves the study to understand the topic. Simple to have domain knowledge.
Data preprocessing	In this stage, the data are identified and selected. It has data sampling, cleansing, and reduction. This stage is necessary to remove the dirty/noise data and outliers to improve quality.y
Data Mining	This process involved selecting machine learning techniques that will be used to create a model and come out with the desired outcome .e
Pattern Interpretation	Focus on checking the performance of the developed model. It is just a model evaluation process.
Model deployment	This stage is for putting the model to use.

2) Data preprocessing: Data preprocessing is a process that involves various techniques like data selection, data cleaning, data integrations, data reduction, and data transformation. For example, the collected dataset from Kilimanjaro, Arusha, and Manyara had two types of data set the first dataset of client information of 6346 samples with 11 features, and the client index information of 7226 samples with 13 features. The features of client information included client_id, Date, Sex, Age, Residence, contact no, CTC_no, Position, Marital status, HIV knowledge, and the number of HIV indexes. The client index features include client_id, client-name, Contact number, CTC _number, Position, Registration type, Date, Site name, Region, Sex of _index, Age, Type of relationship, and HIV status. Based on the nature of the data and literature review, the preprocessing techniques performed as follows:

The selected dataset was cleaned by ignoring features with no value, and the most occurring feature-filled the missing values; the duplicated value was identified and cleared. The dataset used to make predictions was the client information. Later the two datasets were combined for data visualisation.

Data reduction was made on the following features: client id, date, residence, contact number, and CTC number. The removal was due to the following reasons; the features had no impact on the target (client id, contact no, and CTC number). The features had no values (date and residence). Lastly, the data was transformed into a suitable format for model development. The categorical data were converted into 1 and 0, respectively.

3) Data visualization: Data Visualization refers to the graphical presentation of the analysed data so a user can get insight from it and make decisions [22][23]. Data exploration was done using python.

4) Machine learning algorithms: There are different ways of solving ML problems. ML can be divided into three major parties: Supervised, unsupervised, and reinforcement. Each model may apply algorithms based on the dataset and intended results [24]. Machine learning models are designed to classify things, predict outcomes, find patterns and make informed decisions.

Based on this study, three algorithms were selected for performance comparison to determine the best algorithm for predicting the number of HIV Index testing (based on literature). These algorithms were XGBoost, Random Forest (RF), and Neural network. The study considered all the three ML algorithms to select the best performing. Therefore, these algorithms are explained hereunder.

a) XGBoost: XGBoost is an ensemble algorithm based on gradient boosting that has been explained to be an efficient and reliable machine learning technique in solving challenges[25]. It is an open-source library that works best in speed, performance, and parameter setup[26]. XGBoost is used in classification and regression predictive modelling problems. XGBoost denotes the best algorithm for competition on the Kaggle [27].

b) Random Forest: Random forest is an ensemble learning technique that uses a network of decision trees. Breiman proposed it in 2001. It is used for classification and regression[28][29]. The random forest technique combines various randomised decision trees. It is applied in larger-scale problems. Random sampling enhances the depreciation of the overfitting problem [24]. The randomly generated dataset is used to train the dataset for the ensemble decision tree. Each decision tree will determine output. Fig. 1 below shows how a random forest algorithm is formed.

c) Artificial Neural Network: The artificial neural network, usually called a neural network, is defined as an interconnection of nodes called neurons[30]**.** It works like the human brain works. A collection of neurons created and connected together enables them to send messages to each other. The network is requested to solve a problem, which is performed repeatedly. The more connection is strengthened, the more success is achieved, and the reduced failure.

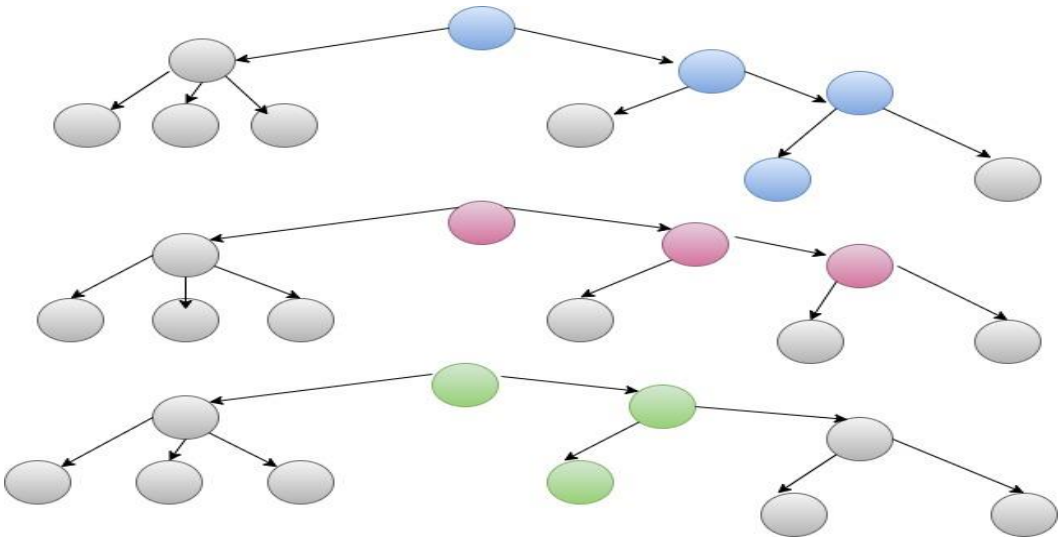


Fig. 1. Random Forests Structure.

The input variables from the data are passed to this neural as a linear connection of various variables. The value multiplied by each characteristic variable is called weight. Now the linear link is applied to nonlinear combinations to provide the ability of nonlinear relationships for neural network modelling. It is used in both classification and regression problems. The artificial neural network is trained by using a random gradient (SGD) and backpropagation algorithm[31]. Fig. 2 shows the structure of an artificial neural network. Each neuro in the input layer represents a column in the input data. Input data is fed to set of neurons and each produces output. Again, each of output is fed to other neuro, which produces another output, which is again fed to the output layer. Error is calculate at this final output layer and again sent back to network for further refine of the output of each neuro. The process is repetitively until the minimal error is obtained.

5) Experimental procedures: The development of models involves major tasks: Acquiring datasets, preprocessing, feature engineering, and model selection. Fig. 3 shows the summary of how the experimental procedures were carried out.

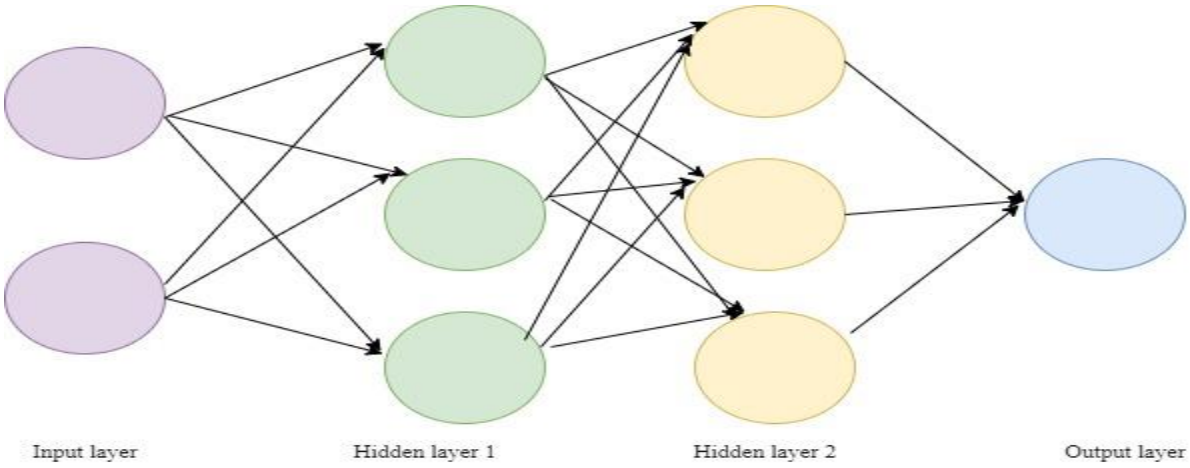


Fig. 2. Neural Network Structure.

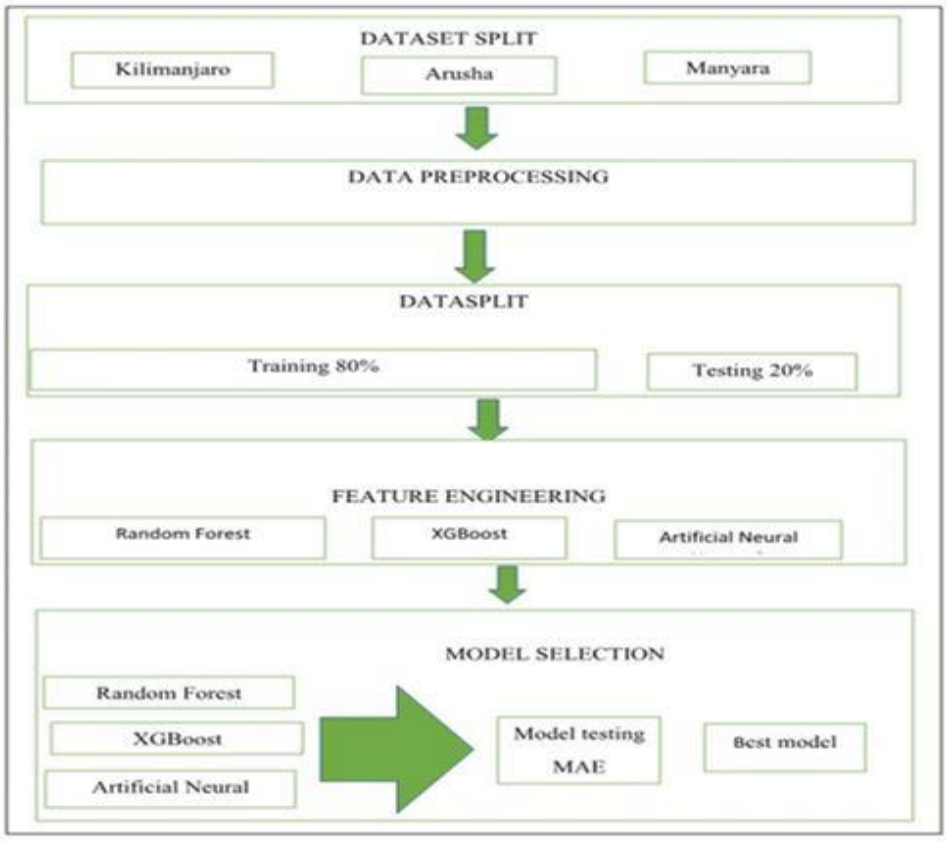


Fig. 3. Experimental Procedures.

6) Evaluation metric: Model evaluation refers to choosing the best-performed model representing data and determining how well the model will work in unseen data. There is a wide variety of evaluation metrics for regression models [32]. According to the literature review, the most used metrics are Mean Squared Error (MSE), RMSE, and MAE. The MSE is calculated as the mean or average squared differences between actual output and predicted target values in a dataset. RMSE is an extension of the mean squared error. MAE score is calculated as the average of the absolute error. In this study, the metric used was MAE due to its simplicity and understandability.

IV. Results and Discussion

A. Results

The subsection explains the results obtained towards developing the HIV index-testing model.

1) Feature engineering: Experiment result from feature engineering showed that people with no knowledge of HIV has a strong coefficient of (0.5). Followed by Marrital_status married (0.175), Age (0.15), Female gender (0.1), Position (influence of someone in the society (0.1), and the rest has a coefficient of less than (0.1). Fig. 4 provides more visualisation of the extracted features using the random forest algorithm. Table II explains in detail the components selected for model development.

2) Data visualization: The section depicts the insight of data from different angles of view. Fig. 5 shows the number of HIV index per client-id. Fig. 6 illustrates the number of HIV indexes by status and site.

Fig. 7 visualise the HIV index versus HIV status and type of relationship. Fig. 8 and Fig. 9 show the number of HIV Index by each region and distributions in term of Age.

3) Model development and evaluation: The result obtained from Model development using three algorithms, as shown in Table III indicates that Random Forest performed well compared to the other two by having the smallest value of MAE: The smaller the value, The desired model.

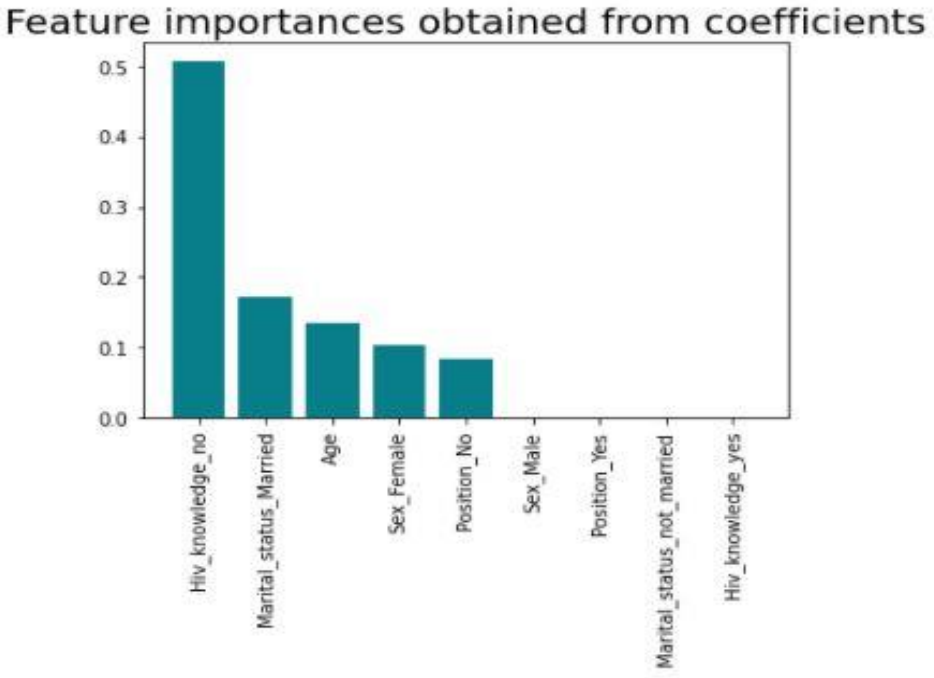


Fig. 4. Feature Engineering.

TABLE II. Selected Feature for Model Development

Variable	Description	Measurement
Age	Age of HI positive client.	number
Sex	Gender of the client (Male/ Female)	1/0
Position	The client is influential in society. Leadership (political, religious, and traditional) values Yes/No	0/1
Marital status	Not married(divorce,widow,widower,never_married)/ married	0/1
HIV_Knowledge	Awareness of client on HIV/AIDS(yes/no)	0/1

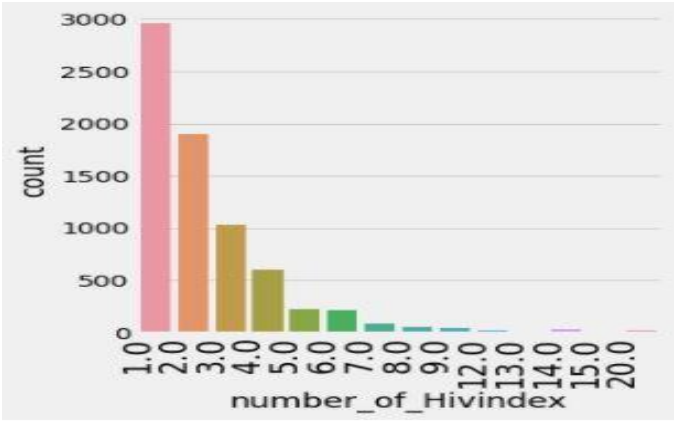


Fig. 5. Number of HIV Index Contacts for each Client_id (Source Google Collab).

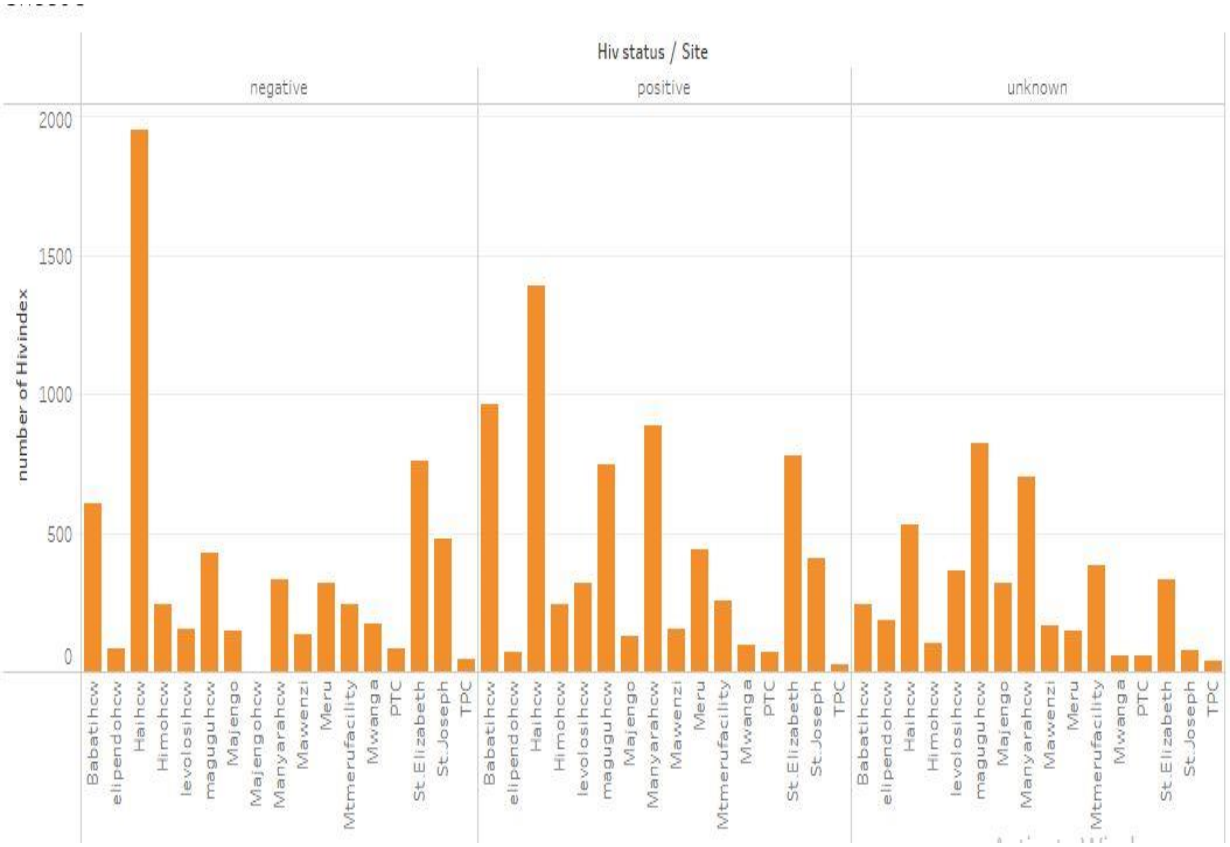


Fig. 6. Number of HIV Index by Status and Site.

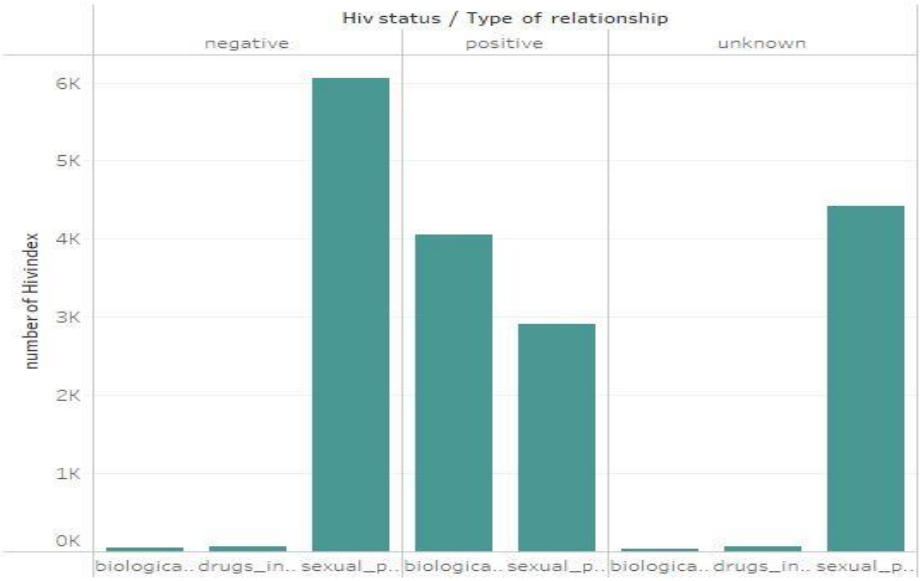


Fig. 7. Number of HIV Index versus HIV Status and Type of Relationship.

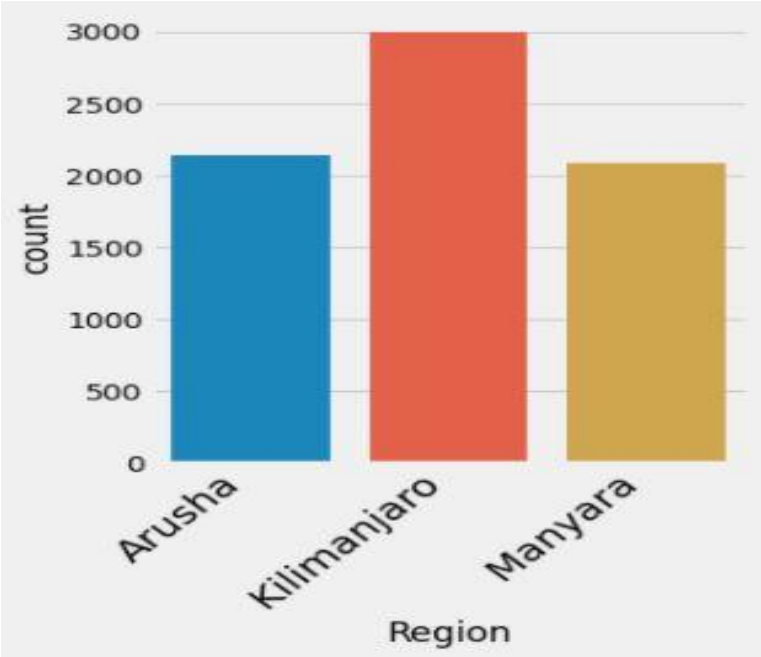


Fig. 8. Diagram shows the Total HIV Indexes in each Region.

number_of_Hivindex distribution across Region by Sex_of_Index

Region ↓	Female	Male	Total number_of_Hivindex
Arusha	81.1%	18.9%	100.0%
Kilimanjaro	83.1%	16.9%	100.0%
Manyara	83.4%	16.6%	100.0%
Grand Total:	**82.6%**	**17.4%**	**100.0%**

Fig. 9. Distribution across the Region.

TABLE III. RESULT OBTAINED DURING MODEL DEVELOPMENT

Serial number(S/N)	Model name	Metric (MAE)
1	Random Forest	1.1261
2	XGBoost	1.2340
3	ANN	1.1268

B. Discussion

The process of understanding the domain knowledge was done thoroughly. Various methods were used to solve the problem to a specified domain. The feature that had a high contribution to the target value was identified. People with no knowledge had led a client to have many client notifications by 35% followed by Position of the client in society 15%, marital status 14%, Age 10%, and Sex 8%.

Data visualisation shows that many clients refer to only one person followed by two and three, while few had up to 12 to 20 people. Kilimanjaro region had high returns compared to the two and a good HIV index specific to the Hai site. The sexual partner notification had a high percentage in information followed by biological children.

The best performance algorithm was a random forest. It had the smallest value of Mean Absolute Error (MAE) of 1.1261. The result remained unchanged after improving the model using the best parameters by GridSearchCV. Lastly, the model was saved ready for deployment.

V. CONCLUSION AND RECOMMENDATION

A. Conclusion

Machine learning is an essential skill in current days. Health care is widely used in many ways, such as decision support, developing medical care guidelines, and applying them in detecting diseases. This paper used machine learning to predict the HIV index and visualisation to help decision-makers develop a suitable intervention strategy to end HIV AIDS as a health threat to society.

However, in achieving the main objective, in addressing the specific goal, the study encountered the following limitations; Missing information in health care data, Lack of enough information in health care such as social-economic and social behaviour information. This information could have an impact on the result. Therefore, the model was developed considering the collected data.

B. Recommendation

Tanzania is one of the sub-Saharan countries with a large rate of people living with HIV. Therefore, client partner notification is vital and can help to yield the target of 95 95 95. However, the study recommends that more researchers and development be required to capture all the required data for better results.

Due to the limitation observed the study recommends that the health care system, especially the unit dealing with HIV/AIDS use the automated system and review the data to be collected for both hospitals and stakeholders to facilitate quality data collection. In addition, HIV knowledge awareness should continuously be given to the community of all ages, and areas.

ACKNOWLEDGMENT

I want to extend my special thanks to the Nelson Mandela African Institution of Science and Technology for granting the opportunity to pursue a master's degree, Center of Excellence in ICT in East Africa (CENIT@EA), for supporting studies and Soft Med company internship took place.

REFERENCES

[1] D.Jerene,*W.Abebe, “Hiv Testing Services Hiv Self-Testing and Partner,” no. December, p. 7, 2017.

[2] U. and C. WH0, PEPFAR, “Partner and Family-Based Index Case Testing,” vol. 1, p. 42, 2015.

[3] M. Katbi et al., “Effect of clients Strategic Index Case Testing on community-based detection of HIV infections (STRICT study).,” Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis., vol. 74, pp. 54–60, Sep. 2018, doi: 10.1016/j.ijid.2018.06.018.

[4] UNAIDS, “Data 2020,” Program. HIV/AIDS, pp. 1–248, 2020, [Online]. Available: https://www.unaids.org/en/resources/documents/2020/unaids-data%0Ahttp://www.unaids.org/sites/default/files/media_asset/20170720_Data_book_2017_en.pdf.

[5] United Nations Joint Programme on HIV/AIDS (UNAIDS), “To help end the AIDS epidemic,” United Nations, p. 40, 2014, [Online]. Available: http://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf.

[6] G. Health, S. Strategies, and W. H. Assembly, “Global Health Sector Strategies 2016-2021 (GHSS) Briefing Note : October 2015,” vol. 2021, no. October 2015, pp. 2016–2021, 2016.

[7] B. Y. Putting, P. At, and T. H. E. Centre, “Prevailing Against Pandemics,” 2020.

[8] National Bureau of Statistics, “Tanzania HIV Impact Survey (THIS) 2016-2017,” Tanzania HIV Impact Surv. 2016-2017, no. December 2017, pp. 2016–2017, 2018, [Online]. Available: https://phia.icap.columbia.edu/wp-content/uploads/2019/06/FINAL_THIS-2016-2017_Final-Report__06.21.19_for-web_TS.pdf.

[9] MOHSS, “National Strategic Framework for HIV and AIDS Response in Namibia 2017/18 to 2021/22,” p. 116, 2017.

[10] SAP Insights, “What Is Machine Learning? | Definition, Types, and Examples | SAP Insights.” 2019, [Online]. Available: https://insights.sap.com/what-is-machine-learning/.

[11] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” BMC Med. Res. Methodol., vol. 19, no. 1, pp. 1–18, 2019, doi: 10.1186/s12874-019-0681-4.

[12] J. Leo, “A reference machine learning model for prediction of cholera epidemics based-on seasonal weather changes linkages in Tanzania.” NM-AIST, 2020.

[13] C. K. Mutai, P. E. McSharry, I. Ngaruye, and E. Musabanganji, “Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa,” BMC Med. Res. Methodol., vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12874-021-01346-2.

[14] Y. Singh, N. Narsai, and M. Mars, “Applying machine learning to predict patient-specific current CD 4 cell count to determine the progression of human immunodeficiency virus (HIV) infection,” African J. Biotechnol., vol. 12, no. 23, 2013.

[15] Q. Zhang, Y. Chai, X. Li, S. D. Young, and J. Zhou, “Using internet search data to predict new HIV diagnoses in China: A modelling study,” BMJ Open, vol. 8, no. 10, 2018, doi: 10.1136/bmjopen-2017-018335.

[16] PEPFAR, HRSA, DIMAGI, and ICAP, “Machine Learning for Predicting Default from HIV Services in Mozambique.”

[17] P. Smyrnov, Y. Sereda, A. Lytvyn, and O. Denisiuk, “Improving HIV case-finding with machine learning ML algorithm has performed better or equally well in comparison with rule based algorithm on making decision who should receive additional recruitment coupons due to higher probability of undiagnosed HIV ca,” p. 1223, 2016.

[18] L. K. Mwango et al., "Index and targeted community-based testing to optimize HIV case finding and ART linkage among men in Zambia," J. Int. AIDS Soc., vol. 23, no. S2, pp. 51–61, 2020, doi: 10.1002/jia2.25520.

[19] N. Mahachi et al., "Sustained high HIV case-finding through index testing and partner notification services: experiences from three provinces in Zimbabwe," J. Int. AIDS Soc., vol. 22, no. S3, pp. 23–30, 2019, doi: 10.1002/jia2.25321.

[20] S. Prabhakaran, "Machine Learning Methods for HIV / AIDS Diagnostics and Therapy Planning," PhD thesis, 2014.

[21] M. J. Pazzani, "Knowledge discovery from data?," IEEE Intelligent Systems and Their Applications, vol. 15, no. 2. pp. 10–12, 2000, doi: 10.1109/5254.850821.

[22] G. Chawla, S. Bamal, and R. Khatana, "Big Data Analytics for Data Visualization: Review of Techniques," Int. J. Comput. Appl., vol. 182, no. 21, pp. 37–40, 2018, doi: 10.5120/ijca2018917977.

[23] Z. M. Khalid, "Big Data Analysis for Data Visualization : A Review," no. January, 2021, dDOI 10.5281/zenodo.4462042.

[24] H. H. Rashidi, N. K. Tran, H. Abb, E. V. Betts, L. P. Howell, and R. Green, "Artificial Intelligence and Machine Learning in Pathology : The Present Landscape of Supervised Methods," vol. 6, 2019, doi: 10.1177/2374289519873088.

[25] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," no. November 2019, 2019, doi: 10.1007/s10462-020-09896-5.

[26] C. Bent and G. Mart, "A Comparative Analysis of XGBoost A Comparative Analysis of XGBoost," no. November 2019, 2020.

[27] "XGBoost for Regression.".

[28] W. Lin, Z. Wu, L. Lin, A. Wen, and J. I. N. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," vol. 5, 2017.

[29] G. Biau and E. Scornet, "A random forest guided tour," TEST, 2016, doi: 10.1007/s11749-016-0481-7.

[30] M. Vakili and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," no. January, pp. 0–13, 2020.

[31] R. Suman, "Understanding Artificial Neural Network With Linear Regression." 2019, [Online]. Available: https://analyticsindiamag.com/ann-with-linear-regression/.

[32] J. Brownlee, "Regression Metrics for Machine Learning," Machine Learning Mastery. 2021, [Online]. Available: https://machinelearningmastery.com/regression-metrics-for-machine-learning/.

Processing of Clinical Notes for Efficient Diagnosis with Dual LSTM

Chandru A. S[1]
Research Scholar, Department of CSE, Visvesvaraya Technological University (VTURRC), Karnataka, India

Seetharam K[2]
Professor, Department of CSE
Jnanavikas Institute of Technology, Bangalore, India

***Abstract*—Clinical records contain patient information such as laboratory values, doctor notes, or medications. However, clinical notes are underutilized because notes are complex, high-dimensional, and sparse. However, these clinical records may play an essential role in modeling clinical decision support systems. The study aimed to develop an effective predictive learning model that can process these sparse data and extract useful information to benefit the clinical decision support system for the effective diagnosis. The proposed system conducts phase-wise data modeling, and suitable text data treatment is carried out for data preparation. The study further utilized the Natural Language Processing (NLP) mechanism where word2vec with Autoencoder is used as a clustering scheme for the topic modeling. Another significant contribution of the proposed work is that a novel approach of learning mechanism is devised by integrating Long Short Term Memory (LSTM) and Convolution Neural Network (CNN) to learn the inter-dependencies of the data sequences to predict diagnosis and patient testimony as output for the clinical decision. The development of the proposed system is carried out using the Python programming language. The study outcome based on the comparative analysis exhibits the effectiveness of the proposed method.**

Keywords*—Clinical notes; natural language processing; diagnosis; long short term memory; convolution neural network; autoencoder

I. Introduction

The diagnosis is a critical part of the healthcare system that decides the kind of treatment that needs to be given to the patient and builds the entire treatment strategy. Initial assessment in the diagnosis is a crucial step, and if it goes wrong, it will lead to lots of consequences. One interesting research study has reported that 65% of the medical mishaps are due to wrong diagnosis only and 11% of these cases result in death [1-2] Therefore, developing intelligent models is essential to reduce false diagnoses and to help experts make the right decisions for patient treatment and well-being. However, building efficient diagnostic systems requires the availability of relevant data and predictive models for real-time deployment. Recent advances in Machine Learning (ML) technologies have brought opportunities to improve healthcare and enhance patient outcomes. Furthermore, clinical records are a decent resource that provides a crucial scheme and scope of optimizing the diagnostic process. The clinical notes are nothing but the text written by the physician and the medical experts who have admitted and treated the patient. These clinical notes usually have two parts to it, namely, i) Patient testimony ii) Doctor's notes. These two contain different types of information, which acts as a powerful resource providing detailed patient conditions and clinical inference, which usually cannot be obtained from the other components of the electronic health record [3-4]. These two parts of the clinical notes have distinct importance in order to make diagnosis better. Based on this clinical information, few studies have shown that among the patients who are entering the hospital from the Emergency Room (ER), 29% of them are unconscious and 40% of them have some type of psychological episodes [5-6]. Among them, the patients who needed Cardiopulmonary Resuscitation (CPR) during the admission to hospital via ER, only 11% survived till the discharge [4]. Among the patients who enter the hospital from Out Patient Department (OPD), it is shown that 80% of the patient (Non-psychological conditions) testimony is fully reliable [7-8]. This proves the point that initial diagnosis is the critical step in the healthcare cycle. Hence depending on the entry point of the patient, the system needs to be designed in such a way that it gives higher importance to patient testimony if the patient is entering from the OPD, and gives lower importance to the same while the patient is entering from ER [9-10]. Recent advances in technologies have shown that natural language processing (NLP) and ML algorithms can be used to build an effective Clinical Decision Support (CDS) system to benefit a successful diagnosis with high scores. However, most of the existing works on CDS have employed standard ML algorithms. On the other hand, the clinical notes contain sparse information, abbreviations, unusual grammatical structures and are high-dimensional. Creating models that learn useful representations of clinical texts is a challenge. While ML algorithms learn without human intervention, preparing the suitable data for ML algorithms needs the right algorithm and tuning it for optimal results. In general, the linguistics of these clinical data requires distinct modeling because they contain a degree of ambiguity that requires the use of different approaches and multiple efforts to come up with the most efficient solutions. Therefore, this paper intends to suggest an effective diagnostic system based on the joint approach of NLP and deep learning techniques. The main contribution of the proposed study is highlighted as follows:

- To emphasize the usage of discharge summary of a patient in order to extract more information data associated with admission of patient for leveraging diagnosis.
- To develop an NLP model for facilitating a unique diagnostic process on the basis of clinical notes of patient which could improve accuracy to next level.

- To deploy Autoencoder for carrying out clustering of text in medical dataset that can classify between doctors notes and clinical testimony of patient.
- Further, novel and efficient Dual-LSTM is built that reveals a different importance to the clinical text depending on the result of the text clustering. Basically, it has two levels of importance to the patient's testimony and Doctor's notes and importance may vary depending on where the patient has entered the hospital from.

The prime motivation of the proposed study is to harness the strength of machine learning approach in a unique fashion in order to carry out reliable diagnostic of the disease. At present, the diagnosis of the disease is mainly carried out considering the medical report of a doctor, whereas various information contextual information could be present. Therefore, this could be further enriched if the information is further provided in the form of patient testimony. Hence, a system is developed in such a way that given the data, of various tests and testimony of both patient and doctors, the system can diagnose the patient.

The remaining part of this paper is structured as follows: Section-II presents a brief review of previous research works; Section III discusses the system design and dataset; Section IV elaborates on the implementation procedure adopted in the proposed system; further result and performance analysis is presented in Section V and finally, Section VI concludes the entire effort and findings.

II. Related Work

This section presents a brief review of literature in context of modeling clinical decision support system based on the machine learning technique and clinical data.

A recent work done by Mustafa and Azghadi [11] provided a detailed review study on applying ML techniques for clinical notes in the healthcare industry. The authors have also discussed potential challenges in working with clinical data and highlight open research issue. The authors have also discussed the concept of AutoML and highlights its benefits for processing clinical notes. However, a data treatment operation is effective in predictive modeling performance, especially when dealing with complicated data like clinical data. The research work in the direction of treating clinical notes is carried out by the Kaur et al. [12] suggested a rule-oriented technique for correcting clinical text data. The authors have applied word correction rule to recognize the term and its definitions. Further supervised ML classifier support vector machine is applied to give suitable treatment for cleaning clinical data. This study has demonstrated effectiveness of combination of the rule-based and ML technique in working with the clinical data. The work or Hassler et al. [13] has shown the importance of the clinical data treatment in the predictive modelling. The first step is data preparation based on the statistical and semantic analysis and new features are then extracted. Further, imputation is done to handle the missing data using ML technique. Another work in the similar research line is done by Kashima et al. [14], where a comprehensive analysis is made regarding the impact of preprocessing at every stage of classification process. The authors have done removal of stop words, lemmatization, normalization and stemming for modelling text data which is then vectorized using Bag of words. Further, a logistic regression approach is used in the classification phase. The study outcome claims that normalization and error correction have a highly positive impact on the classifier's performance. Ferrao et al. [15] provided a roadmap to handle complex medical data using preprocessing techniques. In this study, a phase wise data handling operation is shown that includes error identification, treating missing and redundant data, feature analysis, and information retrieval process. In the work of Mishra and Yadav [16] the preprocessing operation over medical data includes k-means imputation, transformation of discreate value, normalization, random forest-based feature selection. In this study the authors have implemented various ML techniques which achieved higher accuracy with preprocessed data. Once the data are treated and cleaned, their features need to be analyzed and selected. However, the above discussed approaches are specifically focused on the preprocessing operations, and adopted standard approach of feature selection. They have not emphasized effective feature modelling from the perspective of the feature engineering problem, which is an important concern in the predictive modelling, its validation and acceptance in the clinical industry. Although there are few research works in this direction with extensive feature engineering, they need optimization or customization in the design of learning models. Teo et al. [17] attempt to predict the chances of patient readmission in hospital using various ML technique. The outcome illustrates complex deep learning technique outperforms shallow ML models. The work of Spasic and Nenadic [18] gave systematic evidence on the performance of ML model trained on the clinical text. The authors have examined the variety of NLP operations supported by ML techniques. The work of Ye et al. [19] employed unified medical language system and Convolutional Neural Network (CNN) to predict the mortality rate. In this work, the authors have used mimic-III dataset which is applied with concept unique identifiers and entity embedding for the textual feature representation. The usage of unified medical language system with ML-NLP is also seen in the study of Weng et al. [20] for the classification of medical sub-domain. Apart from this, Metathesaurus, are Semantic Network used to extract features which is then combined for the classification process using supervised learning. Kumar et al. [21] developed a classification model to predict whether a morbidity occurs for an individual by analyzing his/her medical records. This study utilized pre-trained word2vec, GloVe, fastText, and sentence encoder embeddings for the classification. Topaz et al. [22] performed comparative analysis between the rule-oriented approaches and NLP-based ML techniques for fall detection from the clinical notes. The work done by Poul et al. [23] has developed a linguistics-driven framework using genetic programming to predict the risk of suicide from the analysis of the clinical data. Using clinical notes, Huang et al. [24] applied bidirectional encoder representations from transformers for forecasting hospital readmission. Prabhakar and Won [25] presented hybrid learning model for medical text data classification. The hybrid model is presented based on hybrid Long Short-Term Memory (LSTM) A bidirectional gated recurrent unit is implemented to reduce the human effort in

modelling data and feature selection. Moqurrab et al. [26] combined CNN, BI-LSTM and discriminant model for the extraction of the clinical entities from the medical notes. Detecting clinical entities accurately can be helpful in maintaining the confidentiality of medical data, which increases trust between users and medical organizations. Table I other significant learning approaches in the context of processing clinical notes for efficient diagnosis.

Hence, there is much research work being carried out in the literature for modelling clinical notes and predictive system for clinical decision support system. Despite numerous works, there is significant issues associated with effective data processing, and predictive modelling. These significant issues are briefly highlighted as follows:

- It has been analyzed that most of the existing approaches have adopted common mechanism for data preprocessing such as normalization and stemming.
- The previous studies have not carried exploratory analysis to understand the nature of dataset and requirement of preprocessing operations.
- Lack of novelty in the design of machine learning model. The existing approaches do not emphasize the modelling of effective learning systems specific to the clinical data.

Therefore, the problem statement for the proposed system can be expressed as "it is quite challenging to design an optimal predictive model along with better treatment operation on the unstructured, and sparse clinical notes for an effective diagnosis support system"

TABLE I. SUMMARY OF THE EXISTING LITERATURE

Citation	Context	Clinical notes processing	Predictive Modelling
[27]	ICD coding	TF-IDF	CNN and Decision Tree
[28]	Detection of anastomosis leakage	BoW	SVM
[29]	Performance of Predictive model	BoW, Word2Vec	Linear regression (LR) and K-nearest neighbor
[30]	Classification of Medical codes	Glove	LR, CNN, LSTM
[31]	Extraction of medication and adverse drug event	Word2Vec	SVM, LSTM, CNN
[32]	Clinical coding analysis	Word2Vec, Glove	SVM, LR, RF, CNN, LSTM
[33]	Classification of diagnosis codes in discharge notes.	Glove	RF, SVM, CNN
[34]	Feature engineering	cTakes	SVM
[35]	Knowledge extraction	cTakes	SVM, KNN
[36]	automated ICD coding	Word2Vec	Deep neural network

III. SYSTEM DESIGN

The development of the system is done using deep learning techniques, which combines mechanism of both LSTM and CNN architecture. The proposed system also makes use of both NLP (Word2Vec) with autoencoder algorithms to make the final diagnosis. The system is set up as a classification learning model where ICD-9 Codes of the diseases are considered as classes. For this purpose, MIMIC-III [9] dataset has been used, which is collected by Beth Israel Deaconess Medical Center. The schematic architecture of the proposed system is shown in figure 1.

The bock diagram of the entire system is as shown in the figure 1. There are two types of data as it can be observed discharge summary and admission type. The admission type is given to train the dual LSTM (LSTM+CNN). Therefore, proposed dual LSTM knows which data it should give importance to depending on ADMISSION_TYPE. The response of this system is diagnosis i.e., the model outputs the diagnosis based on given data. The diagnosis is encoded in the output with one hot encoding technique. Also, a mechanism of NLP i.e., Word2vec is used to identify whether the data is clinical notes or patient testimony, which is achieved by text clustering operation by the proposed dual LSTM. This is discussed in detail in section 3.5. Hence ultimately, the dual-LSTM mentioned over here is most suited for processing medical data which is the novel contribution of this study.

A. Dataset

The MIMIC-III dataset contains total of 26 tables which contains the data of over 40,000 Patients with their personally identifiable information (PII) is deidentified. Hence even though dates provided in the data are wrong, it is made sure that vital data like age of the patient during admission and the number of days stayed in the hospital are not being changed. The deidentification of PII is done in order to protect the patient privacy.

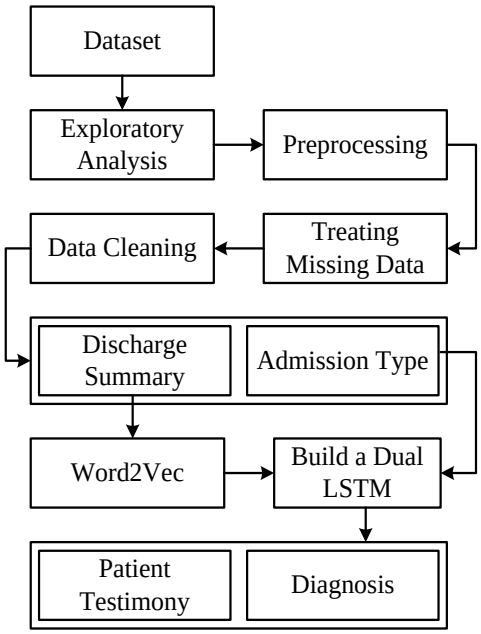


Fig. 1. Block Diagram of the Proposed System.

B. Scope of the Study

This study is limited to non-psychological patients only hence the ICD-9 codes which are between 290-319 are rejected since they represent mental disorder. Even though the dataset contains several other medical records, only discharge summary is being considered.

IV. Methodology Implementation

This section discusses the methodology adopted in the proposed system for the prediction of diagnosis and patient testimony.

A. Exploratory Analysis

In this phase of study, the dataset adopted is analyzed and it is found that it consists of 26 tables. Among the 26 tables, only 4 are important to the present study and they are, i) NOTEEVENTS.csv ii) DIAGNOSES_ICD.csv iii) D_ICD_DIAGNOSIS.csv and iv) ADMISSIONS.csv. However, the in the study not all the columns in these tables are being considered. Only a select few columns which are necessary for this study are being considered here.

NOTEEVENTS (α):

This table contains important notes written by various therapists and nurses during the patients' hospital stay. Following columns are considered in this table.

- SUBJECT_ID: A unique identifier for the patient.
- HADM_ID: A unique ID given to hospital stay. A single SUBJECT_ID has many HADM_IDs. Used as foreign key.
- CHARTTIME: Date and time at which the note was charted.
- CATEGORY: Type of note.
- ISREEOR: If there is an error with the note and needs repetition.
- TEXT: content of the note.

DIAGNOSIS_ICD (β):

This table contains the patient's final diagnosis in form of ICD-9 codes. Following columns are being considered.

- SUBJECT_ID: A unique identifier to each patient.
- HADM_ID: A unique ID given to hospital stay. A single SUBJECT_ID has many HADM_IDs. Used as foreign key.
- ICD9_CODE: Diagnosis made for the admission.

D_DIAGNOSIS_ICD (γ):

This table contains mapping of the ICD 9 code to name of the disease.

- ICD9_CODE: ICD9 code.
- SHORT_TITLE: name of the disease.

ADMISSIONS (θ):

This table consists of the information about the admission of the patient. Following columns are being considered.

- HADM_ID: used as primary key in this case.
- ADMITTIME: time of admission.
- DISCHTIME: time of discharge.
- ADMISSION_TYPE: ER or OPD.
- DEATHTIME: time of death of patient (If died during hospital stay else NaN).

B. Preprocessing

In the preprocessing step all the four tables are initially joined using inner join function. The new table is called as master_data_table (**Ω**) fiven as follows:

$\Omega = \theta[\text{HADM\_ID}] \bowtie \alpha[\text{HADM\_ID}] \bowtie \beta[\text{ICD9\_CODE}] \bowtie \gamma[\text{ICD9\_CODE}]$.

In **Ω**, two columns are then removed since they are primary keys for joining the tables. And they no longer hold any significance. They are, i) SUBJECT_ID and ii) HADM_ID. The ICD9_CODE is truncated to first 3 characters. First 3 characters of the ICD9 code always represents a class of disease rather than the full condition. The algorithm is optimized to recognize the class of disease rather than full condition. For example: ICD code 01166 represents TB pneumonia. Which means fluid collection in lungs due to TB. however, ICD code 01170 represents TB pneumothorax which means damage of lungs due to TB. Any ICD 9 code starting with 011 represent conditions happening due to TB. It is enough for the therapist to know that the patient has TB to start the treatment. ADMITTIME is subtracted from DISCHTIME in order to get LOS (Length of Stay) and then both these columns are dropped. All the rows where ISREEOR is true are dropped A new column called mortality is created which is it is marked as "died" if DEATHTIME is not NaN else marked as "discharged". Only the rows which are marked as marked as "Discharge summary" in CATEGORY are retained and rest of them are dropped.

C. Word2VEC

The cleaning of the text column is done with the help of word2vec algorithm. Initially, to avoid the conflict with cases, the entire text is converted to lower case. Further, in order to avoid the empty words, when there are many whitespaces, they are being replaced with a single whitespace. Fig.2 highlights this process where the discharge summary is considered as an input followed by data cleaning and further Deep Neural Network (DNN) Autoencoder is used for exploring Term Frequency (TF) and Inverse document frequency (IDF). This process yield cluster identity which is further used for performing classification in next step.

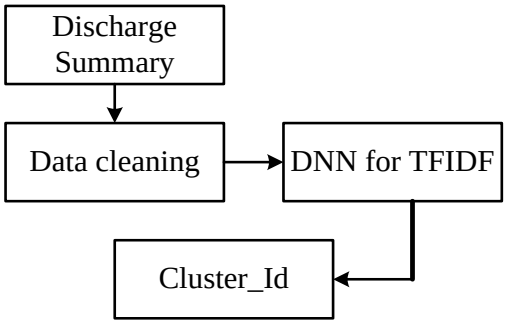


Fig. 2. Block Diagram of the Word2vec Algorithm used in Proposed Work.

As a standard procedure in the NLP, the punctuation marks are removed however the stop words are kept intentionally. Stop words are the words which won't carry lots of meaning ex: in, it, is and etc. They are not removed since they are important while recognizing the person's level of knowledge of medical science. In most of the cases, the patient testimony contains lots of non-technical words whereas clinical notes contain more specific technical words. Once these steps are done, the data is now in a format to be processed by NLP algorithms. Hence, once it is ready, TF-IDF vectorization is applied in the standard method. DNN is trained in such a way that it clusters the data based on the language used. DNN is used to recognize text if it is a patient testimony or clinical notes. The DNN here is setup here in autoencoder configuration. Cluster id 0 represent patient testimony and cluster id 1 represent clinical notes. Here W2V is being trained purely based on the style of writing used in this study: Autoencoder. Autoencoder is a type of DNN which gives exact same output as input. Autoencoder is generally used for data compression. Generally, there are odd numbers of hidden layers in an autoencoder and in this case, there are 5 hidden layers. General applications of AE are, I) Data compression ii) Data de-noising iii) Data generation iv) Data clustering. Data clustering is relatively new technique but not a novel technique. In present study, AE is configured to perform the same.AE is trained with only one class of data. (Either clinical notes or Patient testimony) In this case, AE is trained with Clinical notes. The Clinical notes is text and so is the Patient testimony. When the AE is trained with clinical notes only, then when it is given a Patient testimony as input, it will try to represent that text as a Clinical note. If the input is clinical note, then the output will be same as input However, if input is patient testimony, then output will be totally different compared to input. This is due the fact that AE is trained with only clinical notes. When input is clinical notes, the output is pretty similar. In other words, Euclidian distance between the input and output is less. When input is Patient testimony, output is different compared to input and Euclidian distance is more. We cluster the data based on this Euclidian distance. The threshold for Euclidian distance is set to 0.1. This is done by considering the least loss of the AE.

D. Design of Dual LSTM

This is the most important part of the study where the text from the discharge summary is classified into various diagnosis. This LSTM network contains an attention layer which changes its weights and biases as the admission type varies. The first set of weights and biases are used when the patient is admitted through OPD and second set is used when the patient is admitted through ER. The LSTM gives more importance to patient testimony when the patient is OPD. Depending on the cost center the attention layer allows either first set of weights and biases or 2nd set of weights and biases to process the data. Hence, we will have two specialized neural networks in one. Hence this particular network becomes more robust while diagnosing the patient based on description. Figure 3 shows the architecture of proposed dual LSTM which combines work encoding layer, attention layer, two LSTM layer, one CNN layer and single output layer.

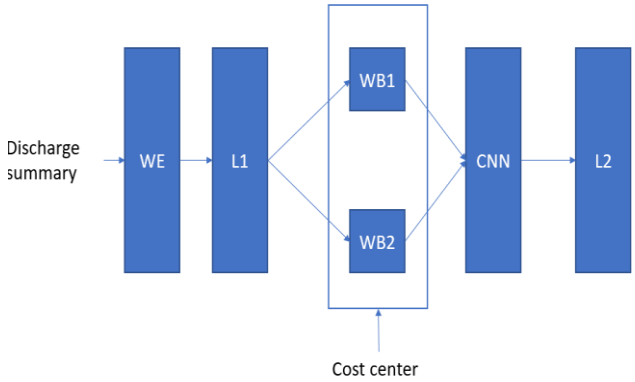


Fig. 3. The Novel Dual LSTM for Processing Medical Data.

V. Result Analysis

The proposed system's design and development is carried out using Python programming language and an anaconda computing environment. This section discusses the outcome and performance analysis of the proposed system.

The above figure 4 shows the analysis regarding number of words per sample versus percentage of sample. IThe graph trend exhibits that the most clinical notes contain more than 1,000 words. This is unsurprising as the documents are discharge summary and they contain all details from admission to discharge.

Figure 5 shows an analysis of the number of words versus a number of documents. The graph trend exhibits the number of documents in which so many words are present and there are rare documents containing 8000 words. However, it can be seen that most documents contain 1000 words.

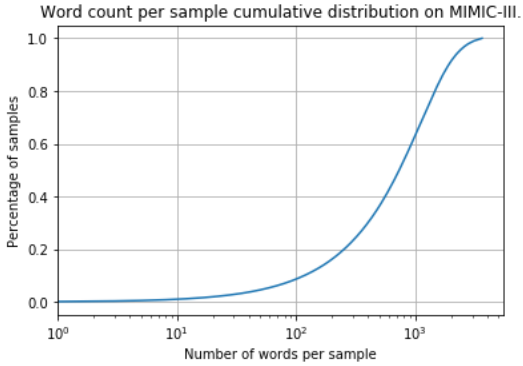


Fig. 4. Word Count per Sample.

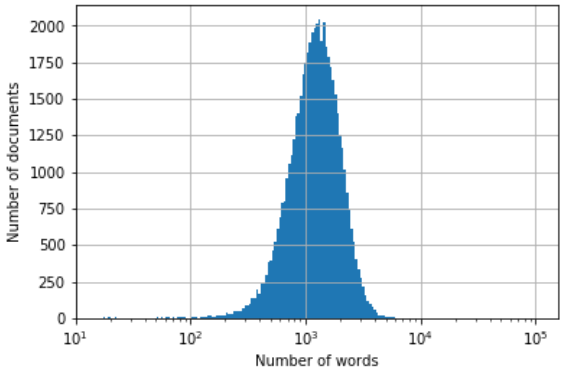


Fig. 5. Number of Words Histogram.

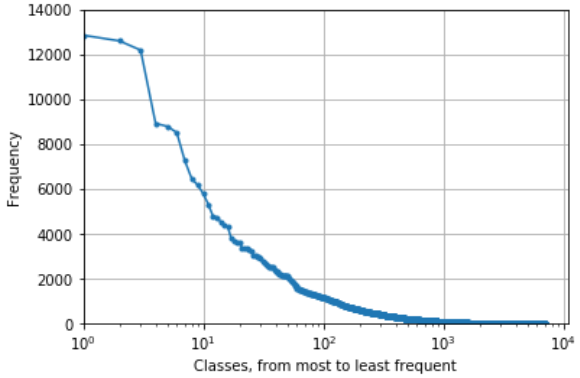


Fig. 7. Classes from Most to Least Frequent.

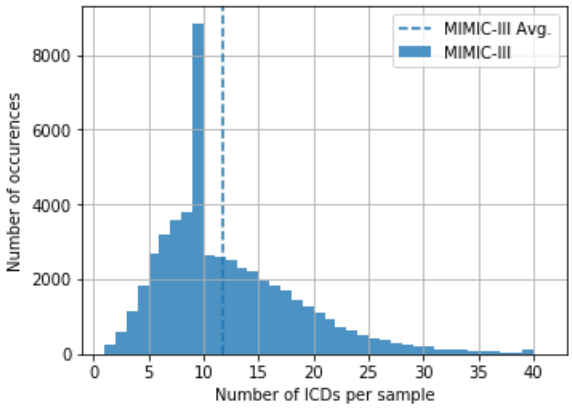


Fig. 6. ICD Histogram.

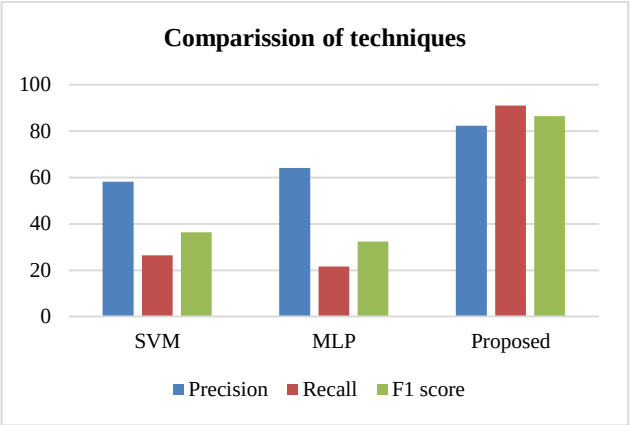


Fig. 8. Compression of Various Techniques.

Figure 6 shows an analysis of the probability distribution of the ICD codes. Based on the graphical trend and frequency of occurrences, data is highly imbalanced, and some ICD codes have been repeated to 8,000 times, whereas others are repeated a very handful number of times. The proposed dual LSTM algorithm handles this imbalance.

The figure 7 shows the analysis regarding the number of classes from most to least occurrences versus frequency. The graph trend exhibits that the less than 100 classes have the highest frequency of occurrences. At the same time, most of the classes have fewer and more frequent occurrences.

The analysis from figure 8 shows that the proposed method shows a better result than all other existing methods. This is due to the fact that we are using LSTM, and text data is a series data. LSTM works best for the series data. The proposed algorithm shows a better result than the other shallow learning and deep learning methods with improvements. As it can be observed, the recall rate has improved greatly. This is because in this study, the more precise and technical clinical notes are given preference; hence, the number of false negatives reduces greatly. As the complete work is carried out on standard MIMIM-III dataset, which is universally approved, the applicability of the proposed scheme suits well with all kind of real-time dataset, which is structured in the form of MIMIC-III dataset or with slightest amendment. It can be used for diagnosis of any form of critical disease using ICD9 codes.

VI. Discussion

The previous section has exhibited the outcome (both individual and comparative) which states proposed scheme excels better performance with respect to existing learning scheme. However, some important learning outcomes are obtained from this which is stated as follows:

- The study considers the performance parameters of precision, recall, and F1-score instead of choosing convention accuracy parameter owing to potential imbalance in classification using multiple clinical attributes.
- Inspite of potential capability to process unstructured data by SVM, still its performance is degraded as it offers challenges in fine tunning hyper parameters. This causes declination of precision (Table II). However, owing to its capability to overcome over-fitting data, its recall rate is better compared to MLP technique.
- Higher Precision score for MLP attributes for its capability to offer robustness against error prone data as well as due to presence of target function, it can offer discrete outcome (Table II). However, MLP is better suited for numerical data whereas the data considered in proposed scheme has both strings and numerical which reduces the recall rate as well as F1-Score.

- Proposed system offers a progressive scheme which uses autoencoders as well as CNN for learning purpose without much reengineering process involved in feature management. This potentially results in improved performance in every aspect (Table II).

TABLE II. COMPARATIVE ANALYSIS

Algorithms	Precision	Recall	F1 score
SVM	58.21	26.36	36.28747
MLP	64.03	21.58	32.28051
Proposed	82.32	90.96	86.4246

Therefore, usage of dual LSTM significantly assisted in overcoming the diagnosis prediction problem that is explored in existing review of literature. However, the primary challenge encountered in the proposed study was to carry out preparation of the data prior to subjecting it to learning operation. This challenge is mitigated by its first module of data preparation where a selected field from the dataset is considered followed by using deep neural network autoencoder.

VII. CONCLUSION

This paper has presented an effective learning system to support the clinical decision process in the patient diagnosis. The proposed system is advanced and highly optimized to process the clinical notes written in rich language. The contribution made in this paper are as follows: i) suitable data treatment and cleaning operation is applied to clinical notes for the processing of NLP and learning mechanisms; ii) Work2vec modeling with Autoencoder is applied to perform clustering of the two distinct clinical classes for the predictive modeling and iii) a dual LSTM is built based on the joint operation of LSTM and CNN deep learning approach. The study outcome exhibited higher performance achieved by the proposed system compared to the shallow machine learning approaches.

REFERENCES

[1] Bhasale A. The wrong diagnosis: identifying causes of potentially adverse events in general practice using incident monitoring. Family Practice. 1998 Aug 1;15(4):308-18.

[2] Rogers WA. Is there a moral duty for doctors to trust patients? Journal of Medical Ethics. 2002 Apr 1;28(2):77-80.

[3] Day SC, Cook EF, Funkenstein H, Goldman L. Evaluation and outcome of emergency room patients with transient loss of consciousness. The American journal of medicine. 1982 Jul 1;73(1):15-23.

[4] Applebaum GE, King JE, Finucane TE. The outcome of CPR initiated in nursing homes. Journal of the American Geriatrics Society. 1990 Mar;38(3):197-200.

[5] Klein MH, Benjamin LS, Rosenfeld R, Treece C, Husted J, Greist JH. The Wisconsin personality disorders inventory: Development, reliability, and validity. Journal of Personality Disorders. 1993 Dec;7(4):285-303.

[6] Putra FB, Yusuf AA, Yulianus H, Pratama YP, Humairra DS, Erifani U, Basuki DK, Sukaridhoto S, Budiarti RP. Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11). In2019 International Electronics Symposium (IES) 2019 Sep 27 (pp. 1-5). IEEE.

[7] Waheeb SA, Ahmed Khan N, Chen B, Shang X. Machine learning based sentiment text classification for evaluating treatment quality of discharge summary. Information. 2020 May;11(5):281.

[8] Diallo B, Hu J, Li T, Khan GA, Liang X, Zhao Y. Deep embedding clustering based on contractive Autoencoder. Neurocomputing. 2021 Apr 14;433:96-107.

[9] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

[10] https://people.cs.pitt.edu/~jlee/papers/cp1_survey_jlee_amuis.pd

[11] Mustafa, Akram, and Mostafa Rahimi Azghadi. "Automated Machine Learning for Healthcare and Clinical Notes Analysis." Computers 10, no. 2 (2021): 24.

[12] Kaur, R. A comparative analysis of selected set of natural language processing (NLP) and machine learning (ML) algorithms for clinical coding using clinical classification standards. Stud. Health Technol. Inform. 2018, 252, 73–79.

[13] Hassler, A., Menasalvas, E., García-García, F. et al. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Mak 19, 33 (2019). https://doi.org/10.1186/s12911-019-0747-6.

[14] Kashina, M., Lenivtceva, I.D. and Kopanitsa, G.D., 2020. Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification. Procedia Computer Science, 178, pp.284-290.

[15] Ferrão, José & Oliveira, Mónica & Janela, Filipe & Martins, Henrique. (2016). Preprocessing structured clinical data for predictive modeling and decision support: A roadmap to tackle the challenges. Applied Clinical Informatics. 7. 1135-1153. 10.4338/ACI-2016-03-SOA-0035.

[16] Misra, Puneet & Yadav, Arun. (2019). Impact of Preprocessing Methods on Healthcare Predictions. SSRN Electronic Journal. 10.2139/ssrn.3349586.

[17] Teo, Kareen & Yong, Ching & Chuah, Joon Huang & Murphy, Belinda & lai, khin wee. (2020). Discovering the Predictive Value of Clinical Notes: Machine Learning Analysis with Text Representation. Journal of Medical Imaging and Health Informatics. 10. 2869-2875. 10.1166/jmihi.2020.3291.

[18] Spasic, Irena, and Goran Nenadic. "Clinical text data in machine learning: systematic review." JMIR medical informatics 8, no. 3 (2020): e17984.

[19] Ye, J., Yao, L., Shen, J. et al. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inform Decis Mak 20, 295 (2020). https://doi.org/10.1186/s12911-020-01318-4.

[20] Weng, WH., Wagholikar, K.B., McCray, A.T. et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 17, 155 (2017). https://doi.org/10.1186/s12911-017-0556-8.

[21] V. Kumar, D. R. Recupero, D. Riboni and R. Helaoui, "Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes," in IEEE Access, vol. 9, pp. 7107-7126, 2021, doi: 10.1109/ACCESS.2020.3043221.

[22] Topaz, Maxim, et al. "Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches." Journal of biomedical informatics 90 (2019): 103103.

[23] Poulin, Chris, et al. "Predicting the risk of suicide by analyzing the text of clinical notes." PloS one 9.1 (2014): e85733.

[24] Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." arXiv preprint arXiv:1904.05342 (2019).

[25] Prabhakar, S. K., & Won, D. O. (2021). Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention. Computational Intelligence and Neuroscience, 2021.

[26] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar and G. Srivastava, "An Accurate Deep Learning Model for Clinical Entity Recognition From Clinical Notes," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 10, pp. 3804-3811, Oct. 2021, doi: 10.1109/JBHI.2021.3099755.

[27] Xu, K.; Lam, M.; Pang, J.; Gao, X.; Band, C.; Mathur, P.; Papay, F.; Khanna, A.K.; Cywinski, J.B.; Maheshwari, K. Multimodal machine

learning for automated ICD coding. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Ann Arbor, MI, USA, 8–10 August 2019; pp. 197–215.

[28] Soguero-Ruiz, C.; Hindberg, K.; Rojo-Álvarez, J.L.; Skrøvseth, S.O.; Godtliebsen, F.; Mortensen, K.; Revhaug, A.; Lindsetmo, R.O.; Augestad, K.M.; Jenssen, R. Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. IEEE J. Biomed. Health Inform. 2014, 20, 1404–1415.

[29] Yogarajan, V.; Montiel, J.; Smith, T.; Pfahringer, B. Seeing The Whole Patient: Using Multi-Label Medical Text Classification Techniques to Enhance Predictions of Medical Codes. arXiv 2020, arXiv:2004.00430.

[30] Karmakar, A. Classifying medical notes into standard disease codes using Machine Learning. arXiv 2018, arXiv:1802.00382

[31] Wei, Q.; Ji, Z.; Li, Z.; Du, J.; Wang, J.; Xu, J.; Xiang, Y.; Tiryaki, F.; Wu, S.; Zhang, Y. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J. Am. Med. Inform. Assoc. 2020, 27, 13–21.

[32] Polignano, M.; Suriano, V.; Lops, P.; de Gemmis, M.; Semeraro, G. A study of Machine Learning models for Clinical Coding of Medical Reports at CodiEsp 2020. In Proceedings of the Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2–25 September 2020.

[33] Lin, C.; Hsu, C.J.; Lou, Y.S.; Yeh, S.J.; Lee, C.C.; Su, S.L.; Chen, H.C. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. J. Med. Internet Res. 2017, 19, e380.

[34] Garla, V.N.; Brandt, C. Ontology-guided feature engineering for clinical text classification. J. Biomed. Inform. 2012, 45, 992–998.

[35] Cobb, R.; Puri, S.; Wang, D.Z.; Baslanti, T.; Bihorac, A. Knowledge extraction and outcome prediction using medical notes. In Proceedings of the ICML Workshop on Role of Machine Learning in Transforming Healthcare, Atlanta, GA, USA, 20–21 June 2013.

[36] Shi, H.; Xie, P.; Hu, Z.; Zhang, M.; Xing, E.P. Towards automated ICD coding using deep learning. arXiv 2017, arXiv:1711.04075.

A Smart Decision Making System for the Selection of Production Parameters using Digital Twin and Ontologies

ABADI Mohammed[1], 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 Chaimae[2]
Laboratory of Mechanics, Mechatronics and Command
Electric Power, Maintenance and Innovation Team, Ecole Nationale Supérieure d'Arts et Métiers (ENSAM-Meknès)
Moulay Ismail University, B.P. 4042, 50000
Meknes, Morocco

ABADI Asmae[3]
INSA, Euro-Mediterranean University of Fez
Fez, Morocco

***Abstract*—Currently, the industrial and economic environment is highly competitive, forcing companies to keep up with technological progress and to be efficient in terms of quality and responsiveness, not only to survive, but also to dominate the market. So, to achieve this goal, companies are always looking to master their production processes, as well as to enlarge their range of products, either by developing new products or by improving old ones. This confronts companies to many problems, including the identification of adequate and optimal production parameters for the development of their products. In this context, a decision making system based on digital twins (DT), case-based reasoning (CBR) and Ontologies is proposed. The originality of this work lies in the fact that it combines three emerging artificial intelligence tools for modeling, reasoning and decision making. Thus, this work proposes a new flexible and automated system that ensures an optimal selection of production parameters for a given complex product. An industrial case of study is developed to illustrate the effectiveness of the proposed approach.**

***Keywords*—Production parameters selection; digital twin; case-based reasoning; ontologies; automation; cyber-physical systems; decision making; artificial intelligence**

I. Introduction

Nowadays, the industrial environment is continuously changing and the industrial competition has become more and more severe due to the consumers exigencies which have become more and more complex and highly personalized. On the one hand, in order to deal with these changes, companies are trying to have flexible manufacturing systems [1] that will allow them to diversify their products and respond to market demands. This diversification requires an efficient choice of the production parameters, in order to realize quality products, with optimal production costs and in the best delays. But in the majority of cases, this choice is made manually, and therefore requires important expenses in terms of time and money. On the other hand, companies tend to make their manufacturing systems intelligent and in real-time [2] to have faster and more efficient production processes. Thus, new concepts have been introduced to the industrial environment namely Digital Twin technologies and manufacturing intelligence [3]. It is in this context that this document has been developed. Indeed, an approach is proposed to automate the choice of production parameters. This approach is resulting from the combination of three artificial intelligence tools, namely:

- Digital Twins (DTs): are an artificial intelligence tool that is capable of copying the operation of production systems in real time and analyzing them, by ensuring a reciprocal interaction between the physical entities and their virtual counterparts [4].

It will allow us to simulate the production process and its parameters that will be automatically computed using the other components of the proposed SPPDT system. The Digital Twin will ensure this simulation in real time and will allow validating the production parameters on the virtual production process before its physical implementation.

- Case-based reasoning (CBR): is one of the tools of artificial intelligence that is based on the use of old functional and efficient solutions to problems encountered in order to treat and solve new similar problems faced [5,6].

The role of CBR in the proposed approach is to ensure a part of the reasoning and decision-making support.

- Ontologies: a set of concepts and parameters that characterize a specialized domain (for example: the pharmaceutical industry) [7], it allows to define the meaning of words (synonyms, thesaurus, ...) and to exchange this information in a unique format/language.

The use of ontologies will ensure interoperability between the different elements of the cyber-physical model of the digital twin. Their use will also ensure the expressiveness of the treated information and preserve their semantics. In addition to that, ontologies will ensure reasoning and decision making for the selection of optimal production parameters.

So, the first section of this paper presents a literature review on the main concepts used for the development of the proposed approach SPPDT (Selection of Production Parameters based on the Digital Twin) namely the digital twins and the ontologies. The second section describes in a general way the proposed approach SPPDT and the functioning of its system. In the following sections, the different modules of the proposed approach are explained in detail, particularly their roles and their functioning. Finally, in the last section, a case study is illustrated in order to prove the good functioning and the efficiency of the developed approach.

II. Related Work

We review in this section the main concepts combined in this work, which are Digital Twins and Ontologies.

A. Digital Twin Concept

The digital twin has become one of the most frequently used tools for managing problems in cyber-physical systems [8]. This concept appeared for the first time at the end of the 1960s, as part of the Nasa Apollo project. This project consisted in the creation of two similar space vehicles: one is sent on a mission and "its twin" is left on earth to follow its state. And so at that time, there was talk about a physical twin that represents the real operating conditions for the simulation of the behavior in real time.

Afterwards, in one of his presentations on product lifecycle management in 2003 at the University of Michigan, Michael Grieves went from this physical model to present a new conceptual model named "mirror space model", and later named "information mirror model" [9]. This one is used to represent virtually and numerically a physical product. Then, in their white paper on the origins of the digital twin, Michael Grieves and John Vickers proposed a general structure for the digital twin that consists of three basic parts [10, 11, 12].

- Physical entity: It usually contains various subsystems that have as role the execution of predefined tasks during operation. In addition to that, these subsystems are equipped with a variety of sensors that collect the necessary information about the working conditions of these subsystems.
- Virtual model: It is a model that reproduces all the characteristics of the physical entity to its geometric and dimensional specifications, its physical properties such as construction materials, the instructions necessary for the correct operation as well as the rules to be applied.
- Connection model: This is the link or interface between the two physical and virtual spaces. It is done through different technologies, including network communication, IoT and network security.

However, according to [11, 13], this structure of the digital twin previously proposed was not complete, and therefore two other dimensions were added to it, namely:

- Services: they can be decomposed into two classes. The first one is a class of business service which schematizes in a simplified and standardized way the inputs and outputs through software interfaces. While the second class is the functional service that ensures the conversion of data and algorithms into support services for the proper functioning of the DT.
- DT data model: It includes the data of all the other elements of the structure, namely the two physical and virtual spaces, the data of the services. In addition to that, it is enriched by the knowledge of the working domain and the merged data of both physical and virtual aspects.

B. Comparing Approaches of Solving Interoperability Problems

According to the ISO-14258-1998 standard, three approaches to achieve interoperability exist, including [14]:

- Integration: It consists in proposing a common standard data model between the different actors of the network. However, the level of compatibility achieved is limited because it is difficult to propose a consensus that is ideally adequate to each of the actors [15].
- Unification: Its principle is based on establishing direct semantic links between the different actors of the network. The problem is that after any modification of a network element, an update must be made at the level of the main model [16].
- Federation: It is based on the idea that each "business" must be able to maintain its own information model to guarantee its meaning and flexibility [15]. Therefore, it is based on logic in order to establish automatic connections between the models used to exchange information between the different actors.

In our context, it is essential to preserve the semantics of the considerable quantity of data that must be exchanged, as well as to ensure a high degree of exchange flexibility manifested in the rapidity and efficiency of decision-making at the right time.

Thus, the federation-based approach is the most suitable to accomplish these objectives, especially since it has become attainable due to information and communication technologies, which propose new modeling paradigms based on the use of ontologies.

C. Overview on Ontologies

According to the ancient Greek, the term “ontology” is composed of two words: “ontos” which means "to be" and the word “logos” which means “discourse”. Then, the definitions of ontology have become various.

In fact, an ontology according to [17], is a characterization that is based on the creation of several axioms that describe the properties of concepts, individuals and relations existing in a

domain of interest, in order to avoid ambiguities of understanding. In the recent decades, Ontologies have been widely exploited in many industrial applications, including knowledge representation [18], and the resolution of interoperability problems [15]. Indeed, the effectiveness of inference Ontologies in handling and solving interoperability problems lies in their ability to [19]:

- Integration and completeness, provided by the expressiveness of the language.
- Embedded intelligence, due to the reasoning capabilities of logical description languages.
- Dynamism and flexibility, due to the queries and web services.

According to [18], these uses of ontologies are possible as a result of its sustained expressiveness and reasoning capabilities.

- Expressiveness: the information treated by ontologies must be expressed with formal languages (a syntax, semantics and rules), in order to be understood by machines. For this reason, the Web Ontology Language (OWL) has been designated as a standard ontology language, recommended by the W3C [20]. Indeed, OWL offers a great capacity to define concepts and relations, based on the description logics (negation, restriction, existence of concepts and even intersection).
- Reasoning: New implicit information can be inferred about instances from the concepts, relations, axioms and rules explicitly defined in the ontology. In fact, ontologies are based on logic and therefore can perform inference operations using inference engines or reasoners.

Thus, in this article, the use of ontologies will play a crucial role, since it will ensure the interoperability between the different elements of the cyber-physical model of the digital Twin, the expressiveness of the treated information and the preservation of their semantics, as well as the reasoning and the help to the decision making.

III. The Global Proposed SPPDT Methodology

The main objective of this article is the automatization of decision making for optimal selection of production parameters. In order to do that, a structured methodology based on Digital Twins (DT), Case Based Reasoning (CBR) and Ontologies is proposed. Indeed, the fusion of these three tools gives birth to SPPDT (selection of production parameters based on the digital twin) methodology.

As shown in Fig. 1, it is composed of two phases: construction phase and operation phase.

The first one (construction phase) contains two modules:

- The Data Collection Module (DCM): This first step is crucial in the proposed methodology. In fact, it is essential to collect the necessary information, either on the product or on the production parameters, which will be useful to make a complete and efficient study.

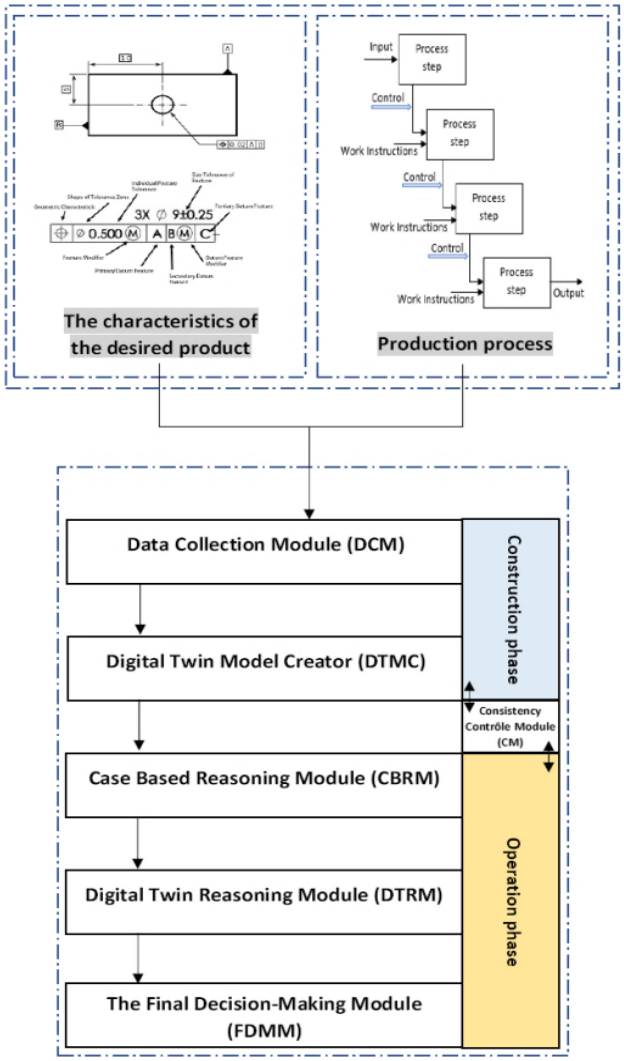


Fig. 1. The Global Structure of the Proposed Automated System for Selection of Production Parameters.

- The Digital Twin Model Creator (DTMC): It is the last module in construction phase and through which the digital twin model of the desired study is established. Indeed, five dimensions are included in this model, namely physical entity (PE), virtual model (VE), connection between PE and VE (CN), DT data (DD) and services (Ss) [13]. The physical entity will be simulated virtually, and the cyber-physical connection will be established based on a manufacturing ontology. In fact, the interoperability between the physical entity and the virtual model will be ensured throughout our DTM-Onto.

While the second phase (operation phase) contains three modules:

- The Case Based Reasoning Module (CBRM): The utility of this module is to understand and solve new problems out of old experiences. To do so, a reasoner recalls and uses a previous case, saved in a database and similar to the actual one to solve this new problem.

Then, the reasoner analyzes the new situation, and tries either to adapt the already existing solutions to the problem, to create an equitable solution to solve it, or to interpret and to critique new solutions.

- The Digital Twin Reasoning Module (DTRM): Before manufacturing a product, the design team requires well determined dimensions and specifications. However, there are always deviations between the design and the manufactured product due to several criteria (production parameters, manufacturing tools, ...). Hence the importance of this module (DTRM) which plays the role of a second reasoner via the manufacturing ontology previously established. Thus, through the formalization of a set of queries, the ontology will provide direct answers to the production team on the production parameters to be configured at the level of physical resources. These queries will be formalized and used to ask the ontology for the optimal values of the production parameters to configure.
- The Final Decision-Making Module (FDMM): redoing an experiment represents a waste of time and money for companies. This is why this module is added to the methodology in order to save and archive the results obtained in previous situations, in data bases, to enrich the manufacturing ontology and to benefit from them in future experiments.
- The Consistency Control module (CCM): This module is responsible for controlling the consistency between the modules of the construction phase and those of the operation phase.

Fig. 2 shows the working process of the proposed SPPDT system.

In fact, the process starts with the collection of necessary information about the production process and the product, including its dimensions and specifications, as well as the number of blocks in the process and their functioning. After that, if the production process is composed of several blocks, it will be divided into many sections; when the production parameters change, the section changes, and then a section may contain one or more blocks. Subsequently, for each section, all possible combinations of production parameters (Xi) will be determined. The Digital Twin of the product is established for the first combination X_1 and the CBR based decision making process is executed. If, the established Digital Twin model does not exist in the case base, the Digital Twin based reasoning process will be executed, otherwise this step will be ignored. In addition, the database is updated by saving and archiving the results obtained and deleting the repeated cases.

All the steps applied for the X_1 combination are repeated for all the other combinations. Finally, the process ends with the determination of the optimal combination of production parameters. The modules of the proposed system will be detailed in the next sections.

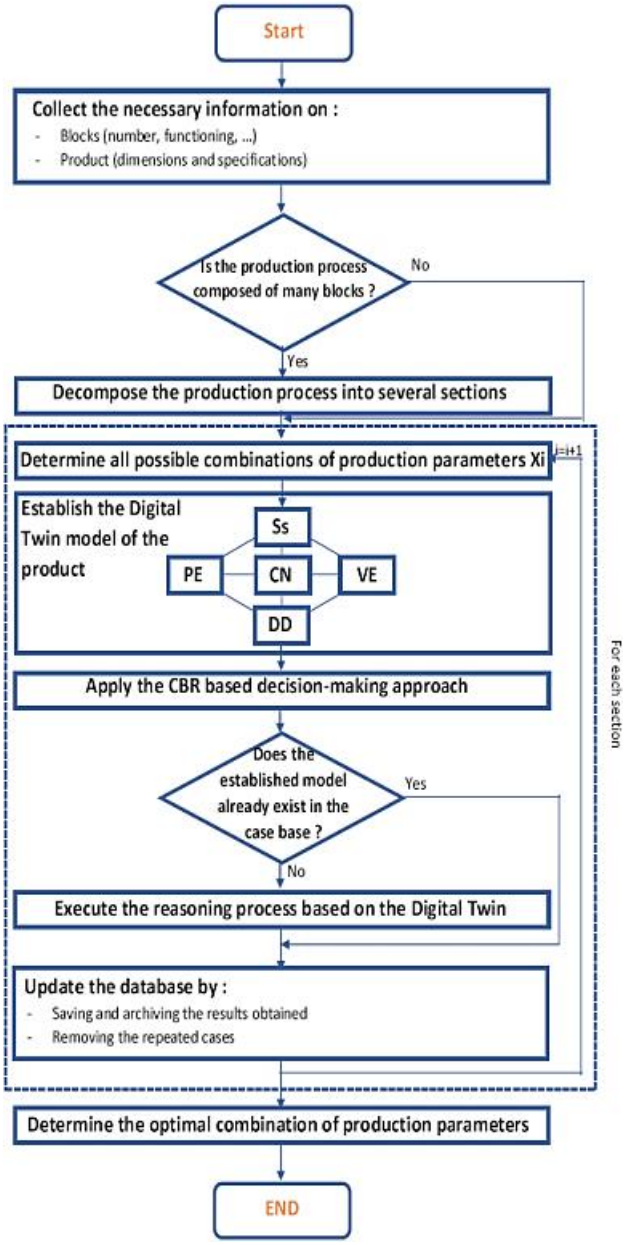


Fig. 2. The Working Process of the SPPDT System.

IV. Construction of the Digital Twin Model

In this section, the digital twin model of the desired study is established by going through several steps:

A. Step 1 : Define the Constraints of the Product

The product constraints are decomposed into [21]:

- Dimensional constraints: which can be used to designate the size of specific entities or the relative location between different entities.
- Geometric constraints: which represent relations between geometric entities such as tangency, collinearity, parallelism, perpendicularity, coincidence of points, symmetry, etc.

In this part, geometric and dimensional constraints are defined using the following notations:

- l: Number of product surfaces;
- m: Number of dimensional constraints in the surface k of the product;
- cd_{kj}: Dimensional constraint, k=1, …, l and j=1, …, m;
- n: Number of geometric constraints in the surface k of the product;
- cg_{kj}: Geometric constraint, k=1, …, l and j=1, …, n.

For dimensional constraints, the measurements cannot be exact in reality, so for each of them a tolerance interval will be defined.

Concerning the geometrical constraints, the types of those existing will be determined for all surfaces (tangency, collinearity, etc.).

B. Step 2 : Determine the Possible Combinations of Production Parameters

The majority of production processes consist of several blocks. These blocks will be grouped into several sections (n sections). Indeed, the section changes when the production parameters change from one block to another.

As shown in Fig. 3, each section contains N parameters to be set (P1, P2, …, PN) and each parameter can take multiple values (A, B, ..., Z).

In addition to that, it should be noted that the production parameters and their numbers can be not the same from one section to another.

For example, P1 of section 1 is not necessarily the same P1 of section 2.

Thereafter, α_i=A*B*…*Z combinations of production parameters can be generated for the i^{th} section, and these combinations are denoted X_{ij}.

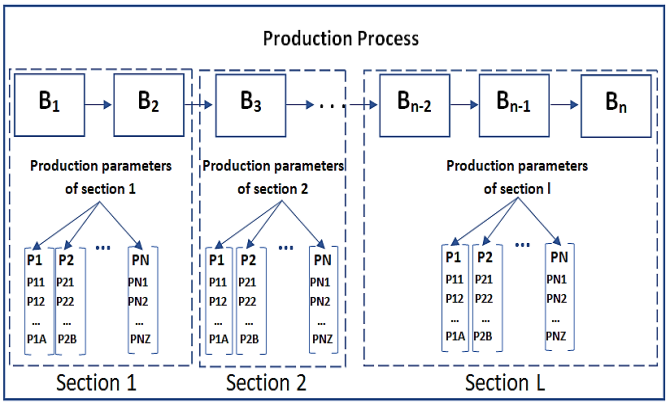


Fig. 3. The Notation used for Indexing Production Parameters.

- With: i=1, …, L and j=1, …, α
- For example, X_{11}= {P11, P21, …, PN1}.

C. Step 3 : Define the Digital Twin Model

Once the product constraints and production parameters have been determined, all that remains is to establish the Digital Twin model of the production process. In fact, the decomposition of the whole process in multiple sections imposes the establishment of several sub-models, which will be gathered later, in order to obtain the global model of the Digital Twin.

The Digital Twin sub-model of each section will be represented as follows:

$$DT_i = \{PE_i, VE_i, Ss_i, DD_i, CN_i\} \quad (1)$$

With:

i=1, …, L;

L is the number of the sections.

And then, the global model of the Digital Twin of the whole process will be as follows:

$$DT = \bigcup_{i=1}^{L} DT_i = \bigcup_{i=1}^{L} \{PE_i, VE_i, Ss_i, DD_i, CN_i\} \quad (2)$$

It should be noted that the connection interface between the physical entity and the virtual model, as well as the Digital Twin Data Model and services are established based on an manufacturing domain ontology created and named "Digital Twin Manufacturing Ontology" (DTM-Onto). This ontology will be developed more in the part of the DTR module.

In this section, the DTM-Onto is constructed for two reasons. On one hand, this ontology will be the connection interface between the physical entity and the virtual model, as well as the data model and the Digital Twin services. On the other hand, the DTM-Onto will be used later to do the reasoning at the Levels of the CBR and DTR modules.

In this article, the DTM-Onto is developed in the ontology editor "Protégé". It is composed of three main elements which are [22,23]:

- Classes: are a set of individuals that describe concepts in a specific domain. In this paper, the classes are related to the elements in the manufacturing domain;
- Object properties: They identify the links between the classes and the individuals;
- Data properties: They define modifiers for ontology classes or establish characteristics of the instances.

Fig. 4 represents the different classes and Object Properties configured on the constructed DTM-Onto. Data Properties will be configured at the case study level.

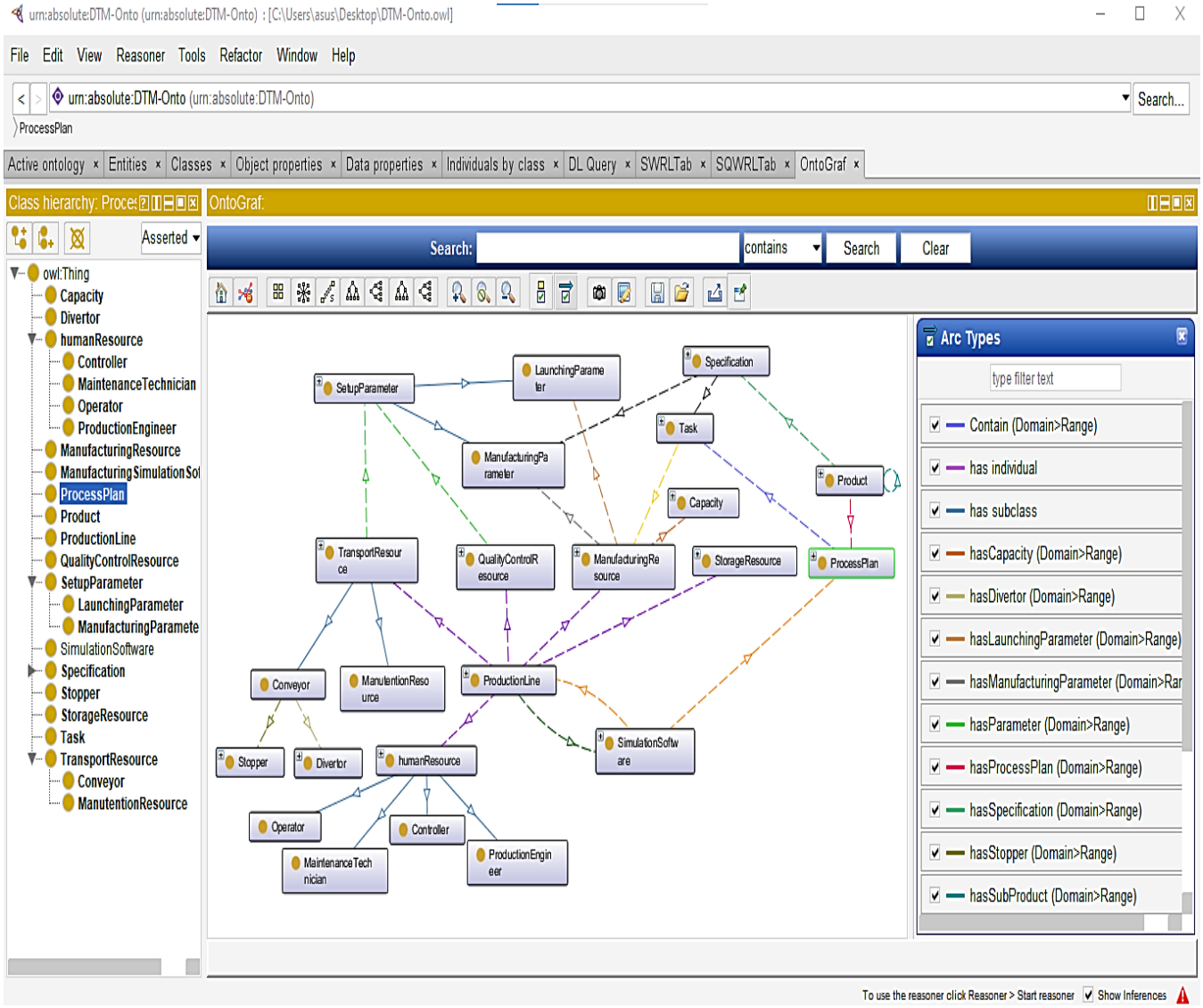


Fig. 4. The General Conceptual Model of the DTM-Onto Ontology Proposed.

V. Description of the Case-based Reasoning (CBR) Module for Optimal Choice of Production Parameters

The case-based reasoning module (CBR module) is based on the use of previous studies of production parameters choices, saved in databases in the company's information system. Thus, the old studies are used to benefit from their results for an optimal choice of production parameters for the new studied case.

As shown in Fig. 5, the proposed working process structure for the CBR module contains three main phases:

A. Preliminary Phase

First of all, an attribution of indexes to the cases is required to facilitate their retrieval. In fact, assume that the case base (CB) contains multiple problems (C_i) which represent the specifications of the desired product, and their results which represent the appropriate production parameters for each section to realize it. Then, the case base can be represented as follows:

$$CB = \{C_1, C_2, \ldots, C_r\} \quad (3)$$

With:

r is the number of problems solved and stored in the case base.

Moreover, each case (C_i) contains L sub-cases (C_{ij}). Each of these sub-cases is composed of two parts: the first one contains the data (production parameters) and the second one contains the generated results (product specifications):

$$\boldsymbol{C_{ij}}: \quad X_{ij} \quad \text{Results} \quad \sum cd_{ij} + \sum cg_{kj} \quad (4)$$

B. Matching Phase

In the matching phase, a comparison between what is desired and what already exists on the basis of cases will be done. In other words, a similarity index between the specifications of the products saved in the case base and those desired in the studied one will be calculated for each section.

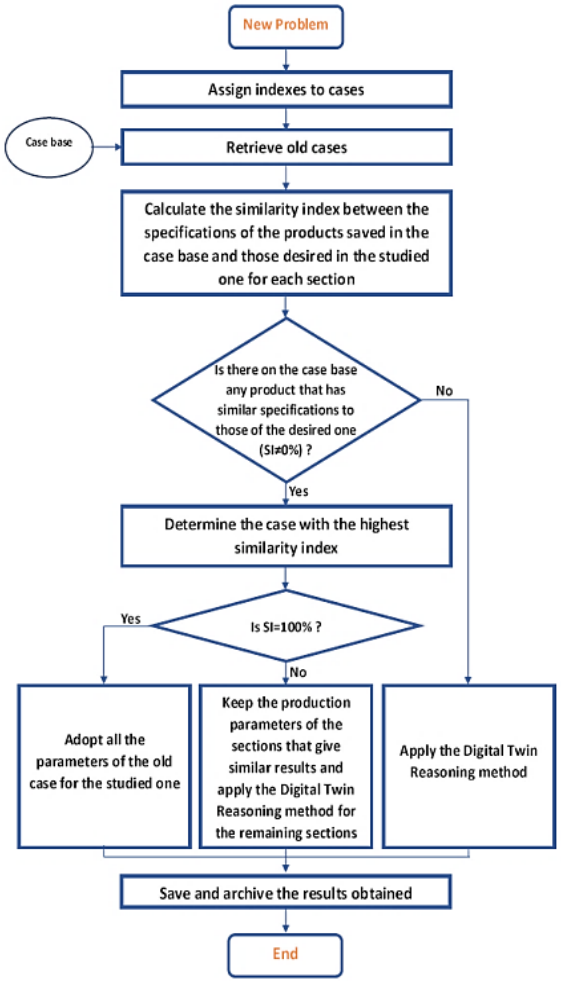


Fig. 5. The Proposed Working Process Structure for the CBR Module.

To do that, the working process shown in Fig. 6 will be executed, with:

- SI: the similarity index (%);
- Cd_i : the set of dimensional constraints existing at the section i;
- $\widetilde{Cd_i}$: the set of desired dimensional constraints in section i;
- Cg_i : the set of geometrical constraints existing at the section i;
- $\widetilde{Cg_i}$: The set of desired geometrical constraints in section i.

At the beginning of the process, "a" is equal to 0 and "i" is equal to 1.

It is clear that in each section a set of dimensional and geometric constraints of the product is realized. So at the beginning, a comparison at the level of section 1 will be made between the dimensional and geometrical constraints wanted on the new product and each old case existing in relation with Section 1: comparison between $\widetilde{Cd_1}$ and each Cd_1 as well as $\widetilde{Cg_1}$ and each Cg_1.

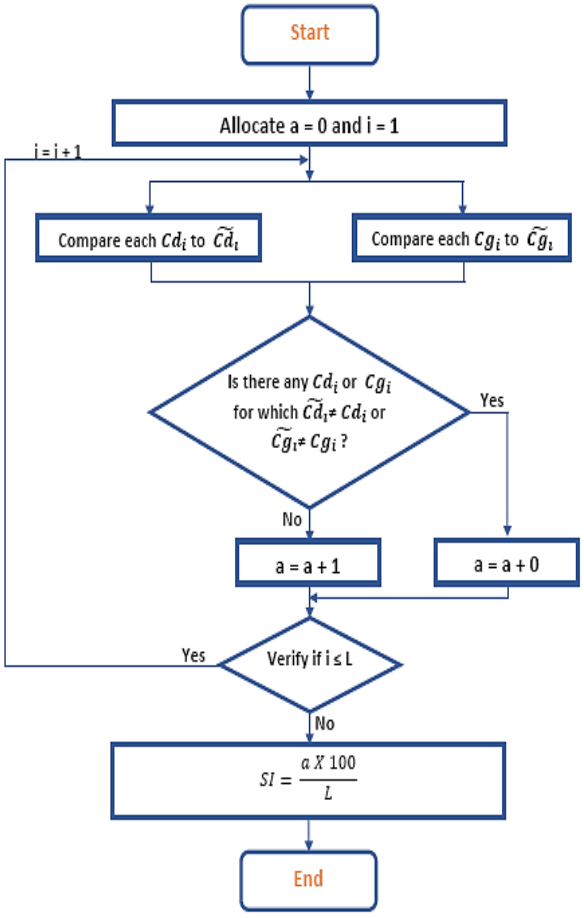


Fig. 6. The Algorithm for Calculating the Similarity Index.

After that, two alternatives can be considered:

- If there is an old product with the same geometrical and dimensional specifications as the new product, a will be incremented by 1.
- If the new value of a is strictly superior to L (the total number of sections of the production process), the comparison cycle is interrupted immediately;
- If not, the comparison cycle will be executed for the next section.
- Otherwise, the comparison cycle is interrupted immediately.

Once the comparison cycle is completed, the similarity index is calculated by the following relation and the process is stopped:

$$SI = \frac{a\ X\ 100}{L} \quad (5)$$

C. Decision-Making Phase

This is the most important phase of the CBR process.

Indeed, after the calculation of the similarity index in the matching phase, three cases are supposed to have:

- SI = 0% : the Digital Twin Reasoning method will be applied;
- 0 % < SI < 100%: the production parameters of the sections that give similar results will be kept and the Digital Twin Reasoning method will be applied for the remaining sections;
- SI = 100 %: all the parameters of the different sections of the previous similar cases will be adopted for the studied one.

Finally, the new results obtained will be saved and archived in the case database for a probable benefit in the realization of new products.

VI. Description of the Digital Twin Reasoning Module (DTRM) for Optimal Choice of Production Parameters

The Digital Twin Reasoning Module (DTRM) is very important in complementing the proposed SPPDT methodology.

The backbone of this module consists of the manufacturing domain ontology (DTM-Onto, which plays the role of connection interface between the physical entities and the virtual entities of the Digital Twin developed before. And due to its reasoning capacities, the DTM-Onto will be used to automate the calculation of production parameters as it will be described later.

Fig. 7 shows the operating system of the DTR module.

The first step will be to formalize the business rules related to each manufacturing process using the SWRL (Semantic Web Rule Language) and to introduce them in the DTM-Onto ontology. For example, for the stamping process, the formalized business rules will allow the calculation of the falling mass of the press to be used, the number of punches required to obtain all the geometric details of the product, the speed of the punches, the punching force, etc. For the machining process, the formalized business rules will allow the calculation of the feed rates, the number of passes, etc.

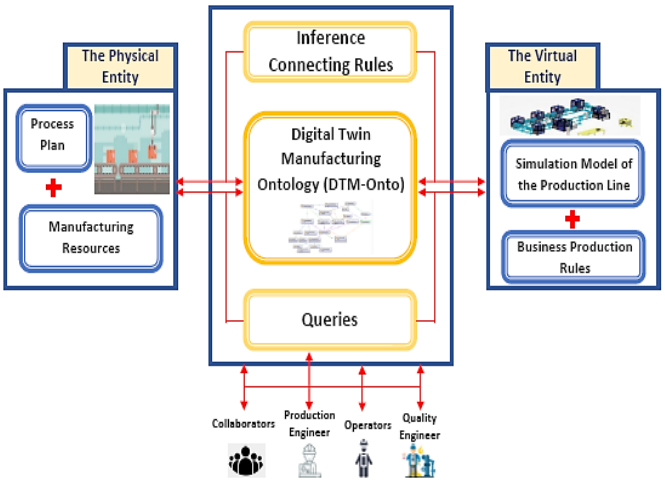


Fig. 7. The Operating System of the DTR Module.

Simulations using flow modeling software such as WITNESS, 3DEXPERIENCE, etc., will be used to validate the production parameters to be used and to virtually visualize the production sequences. Conveyor speeds as well as Pick and Place programs for handling robots will also be determined during this phase.

Once all the production parameters are validated by the simulation, the DTM-Onto ontology will be enriched with them. Indeed, as it will be described in the following, instances and object properties will be added to DTM-Onto to do so.

A second category of inference rules will then be executed. These are the matching rules between the production parameters of the virtual entities and those of the physical entities. These rules will ensure the interoperability at the virtual/physical interface of the DT in a dynamic way.

Finally, through the formalization of a set of queries, the ontology will provide direct answers to the production team on the production parameters to be configured at the physical resource level. These queries will be formalized using the SQWRL (Semantic Query-Enhanced Web Rule Language) and used to ask the ontology about the appropriate production parameters.

VII. Case Study

In this section, an industrial case study is presented to validate the operation and applicability of the proposed SPPDT methodology.

Indeed, the various phases of the proposed methodology are applied, in this section, on the production process of yogurt in a company specialized in manufacturing dairy products.

Initially, the company manufactures small yogurt cups with the geometric and dimensional specifications shown in Fig. 8 and Table I.

To do so, the realization of the final product requires two main parts, namely: a PROCESS part and a conditioning part, but the case study will only focus on the conditioning part.

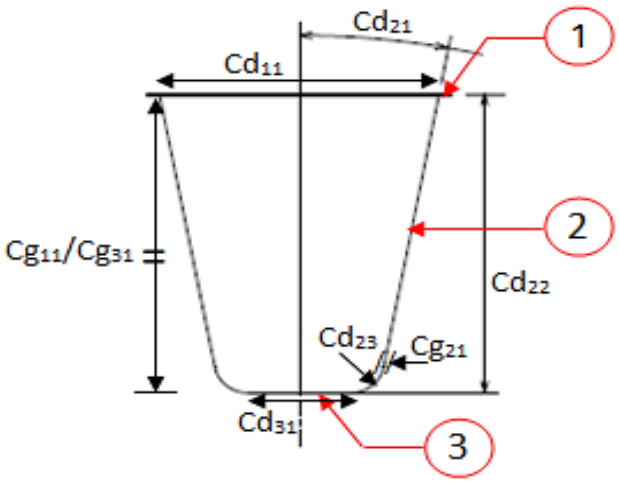


Fig. 8. Indexing of Geometrical and Dimensional Constraints of the Product.

TABLE I. THE DIMENSIONAL AND GEOMETRICAL CONSTRAINTS OF THE TWO CUP SIZES

Dimensional and geometrical constraints	Small cups	Large cups
Cd_{11}	70mm	80mm
Cd_{21}	10°	10°
Cd_{22}	70mm	95mm
Cd_{23}	10mm	10mm
Cd_{31}	10.5mm	10.5mm
Cg_{21}	Tangency	Tangency
Cg_{22}	Symmetry	Symmetry
Cg_{11}/Cg_{31}	Parallelism	Parallelism

This phase is realized on a production line which is composed of several blocks that allow executing a set of operations.

In fact, after unrolling the plastic strip (PS), a heating system consisting mainly of heating resistances and temperature probes (for continuous regulation) allows the heating of the PS edges. This operation facilitates its pecking and thus its transport throughout the conditioning process. This transfer is carried out by means of a pimple chain.

Afterwards, an ionizing deduster removes any foreign matter from the PS.

Immediately after, a heating box driven by a cam press is installed. Its role is to heat the PS surfaces which will undergo a deformation. The next step is to form the yogurt pots in the form of packs of 24 pots (the plastic forming block) and to dose them with a piston doser. At the same time, a polymix unwinder allows, as its name indicates, to unroll the polymix with the help of an automatic splicing system. This system allows changing the reel of the polymix automatically. In turn, the polymix passes through a tunnel of UV lamps allowing its ionization so that a dating can take place afterwards.

Before the yogurt pots are cut into packs (24 pots), the polymix is welded to the already dosed pots in a welding block which is also driven by a cam press.

Once the packs are cut, they are packed in plastic boxes. Their transfer from the sealing block to the case packer is done using a conveyor called pilger conveyor.

To finish the conditioning of the yogurt, the full cases are palletized manually on wooden pallets and then stored in cold rooms.

After a study of the market, the company decided to start manufacturing large size cups. These cups are different from the small ones by their dimensional constraints as it is represented in Fig. 8.

This change of series had a big problem of pots piercing during their production. After the analysis of the major causes of the appearance of this defect, it was shown that the problem comes from the bad adjustment of the parameters of production.

Fig. 9 shows the problem of piercing cups.

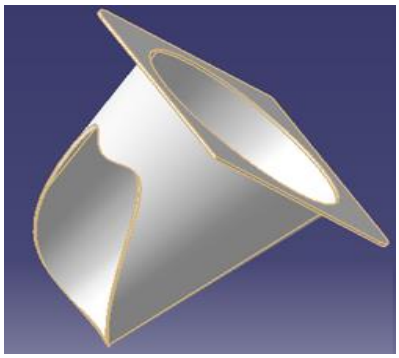


Fig. 9. The Problem of Cups Piercing.

We remind that there are three categories of production parameters to consider. The first one concerns the manufacturing process itself (for the forming of the cups: the heating temperature, the pressure, the depth of the punches,...). The second category concerns the product specifications (dimensional and geometrical constraints of the cups). While the third class concerns food safety (sterilization time, sterilization temperature,...).

This last class of parameters has not been taken into account in the case of study because it respects the food safety standards, and therefore there are no parameters to choose because they are already imposed by the standards.

So, to solve this parameters adjustment problem, the SPPDT approach is executed.

The first step is to place several sensors on the production process in order to copy virtually the physical state of its components.

Fig. 10 shows an extract from the virtual representation of the production process.

In addition to that, the DTM-Onto which plays the role of interoperability interface between the physical entity and its virtual model is enriched with the necessary data for its efficient functioning, namely: the product specifications, the different blocks of the process, the manufacturing parameters, etc.

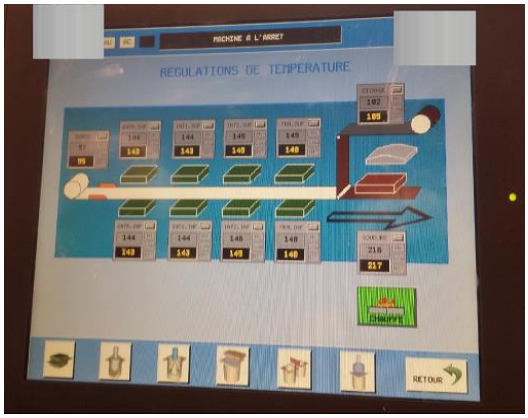


Fig. 10. An Extract from the Virtual Representation of the Production Process.

In the following, the rest of the SPPDT approach is executed on the production process block by block.

For the plastic unrolling block, the first step is to execute the CBR. To do so, the algorithm for calculating the similarity index is applied to the initial product (the raw material: the plastic strip). In our case, the specifications of the plastic strip (material, length, width) used for the manufacture of the large cups are the same as the small ones, the difference that exists is only its thickness, and then: 0 %< SI<100%. This result sends us directly to the use of the Digital Twin Reasoning Module.

This step consists in formalizing a SWRL rule of conservation of the volume between the pot and its raw material. So:

Volume (cup) = Volume (PS)

=> Volume (cup) = Length (PS)*Width (PS)*Thickness (PS)

$$\Rightarrow \text{Thickness (PS)} = \frac{\text{Volume (cup)}}{\text{Length (PS)} * \text{Width (PS)}}$$

And thus the formalized rule is the rule S3 of the Table II.

Fig. 11 shows the plastic strip thickness calculation SWRL rule that we encoded in the SWRL tab of Protégé5.0, its results and their explication that we generated after running the reasoning with the Pellet reasoner.

The same steps applied on the unrolling block are applied on the heating box. The plastic strip introduced into this block has the same specifications (material, length and width) as the one used for the manufacture of small cups, except its thickness which changes, and it is heated to a temperature between the fusion limit and the elasticity limit of the polymer. Consequently: 0 %< SI<100% and the transition to the DTR module is crucial.

This heating temperature is calculated empirically by the following formula:

$$T^{\circ}_{heating} = T^{\circ}_{fusion} - 0.7 * (T^{\circ}_{fusion} - T^{\circ}_{elasticity})$$

This rule is formalized in SWRL by the rule S2 in the Table II, at the DTM-Onto level to automate its calculation. Fig. 12 shows the heating temperature calculation rule encoded in Protégé 5.0, its results and their explanation.

After this, the CBR algorithm is executed on the plastic forming block. In fact, there have been considerable changes in the dimensional and geometric specifications of the new cups that will be manufactured, and subsequently: 0 %< SI<100% and the transition to DTR is mandatory.

What validates the proposed approach is that empirically, when the same production parameters (notably the same punch and the same depth of pass) are kept for the two types of pots, certain non-conformities appear in the product. So to solve this problem and automate the selection of optimal production parameters to use, the DTR is executed. Indeed, three main categories of business rules are defined in the SWRL tab:

- Category 1 (R1-R5 and R9): Formalizes the correspondences between the dimensional specifications of the punches used and the cups formed.
- Category 2 (R6-R8): Formalizes the correspondence between the geometric specifications of the punches used and the cups formed.
- Category 3 (S1): represents the rule of calculation of the depth of descent of the punches. The descent speed remains the same.

TABLE II. BUSINESS RULES PROGRAMMED ON THE ONTOLOGY

Business rules
R1: Product (?Y) ^ hasProductDimensionalConstraint_Cd11(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchDimensionalConstraint_Cd11(?P, ?a)
R2 : Product(?Y) ^ hasProductDimensionalConstraint_Cd21(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchDimensionalConstraint_Cd21(?P, ?a)
R3 : Product(?Y) ^ hasProductDimensionalConstraint_Cd22(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchDimensionalConstraint_Cd22(?P, ?a)
R4 : Product(?Y) ^ hasProductDimensionalConstraint_Cd23(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchDimensionalConstraint_Cd23(?P, ?a)
R5 : Product(?Y) ^ hasProductDimensionalConstraint_Cd23(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchDimensionalConstraint_Cd23(?P, ?a)
R6 : Product(?Y) ^ hasProductGeometricalConstraint_Cg11(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchGeometricalConstraint_Cg11(?P, ?a)
R7 : Product(?Y) ^ hasProductGeometricalConstraint_Cg21(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchGeometricalConstraint_Cg21(?P, ?a)
R8 : Product(?Y) ^ hasProductGeometricalConstraint_Cg22(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchGeometricalConstraint_Cg22(?P, ?a)
R9 : Product(?Y) ^ hasProductDimensionalConstraint_Cd31(?Y, ?a) ^ isAssociatedTo_Punch(?Y, ?P) -> hasPunchDimensionalConstraint_Cd31(?P, ?a)
S1 : Product(?q) ^ hasProductDimensionalConstraint_Cd22(?q, ?i) ^ Punch(?P)^isAssociatedTo_Punch(?q, ?P) -> hasPassingDepth(?P, ?i)
S2 : Product(?q) ^ hasRawMaterial(?q, ?M) ^ hasFusionTemperature(?M, ?f) ^ hasElasticLimitTemperature(?M, ?e) ^ swrlb:subtract(?a, ?f, ?e) ^ swrlb:multiply(?b, ?a, -0.7) ^ swrlb:add(?c, ?b, ?f) -> hasFormingTemperature(?q, ?c)
S3 : Container(?q)^ RawMaterial(?p)^ isConstructedFrom_Plate(?q,?p)^ hasVolume_in_mm3 (?q,?V)^ hasLenght_in_mm (?p,?j)^hasWidth_in_mm (?p,?w)^ swrlb:multiply(?h,?j,?w)^swrlb:divide(?i,?V,?h) -> hasThickness_in_mm (?q,?i)

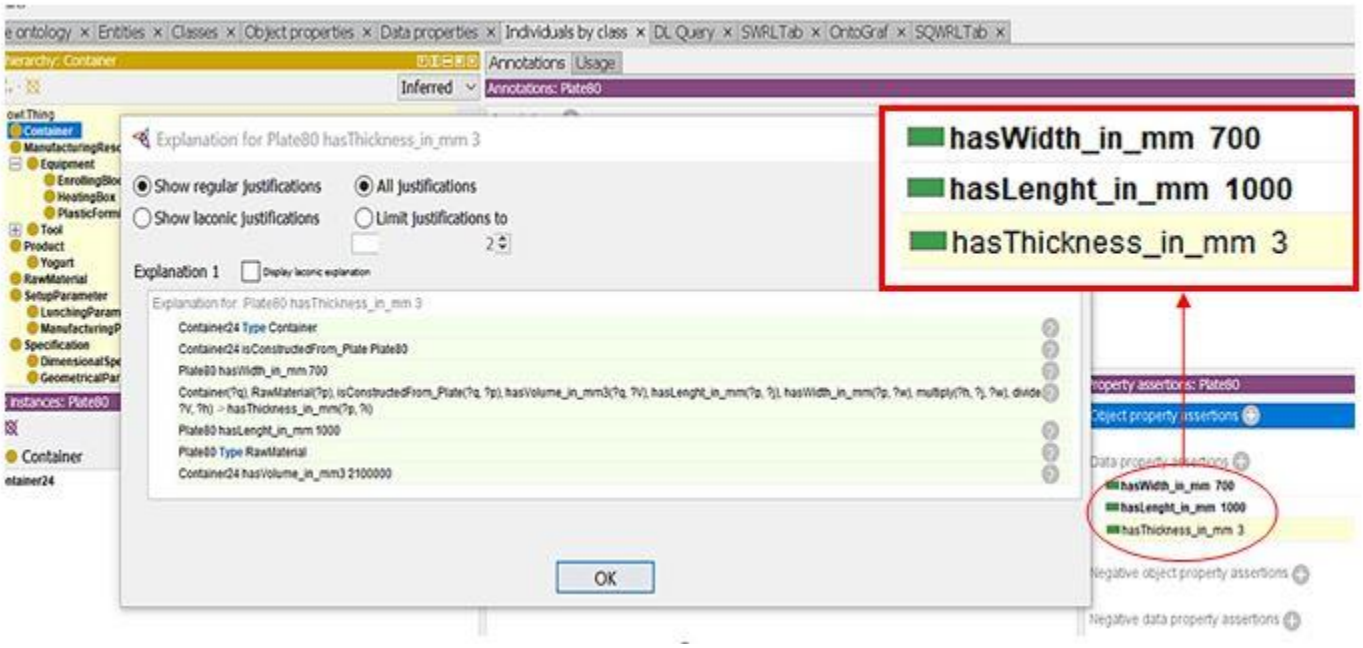


Fig. 11. The Plastic Strip Thickness Calculation Rule Encoded in Protégé 5.0, its Results and their Explanation.

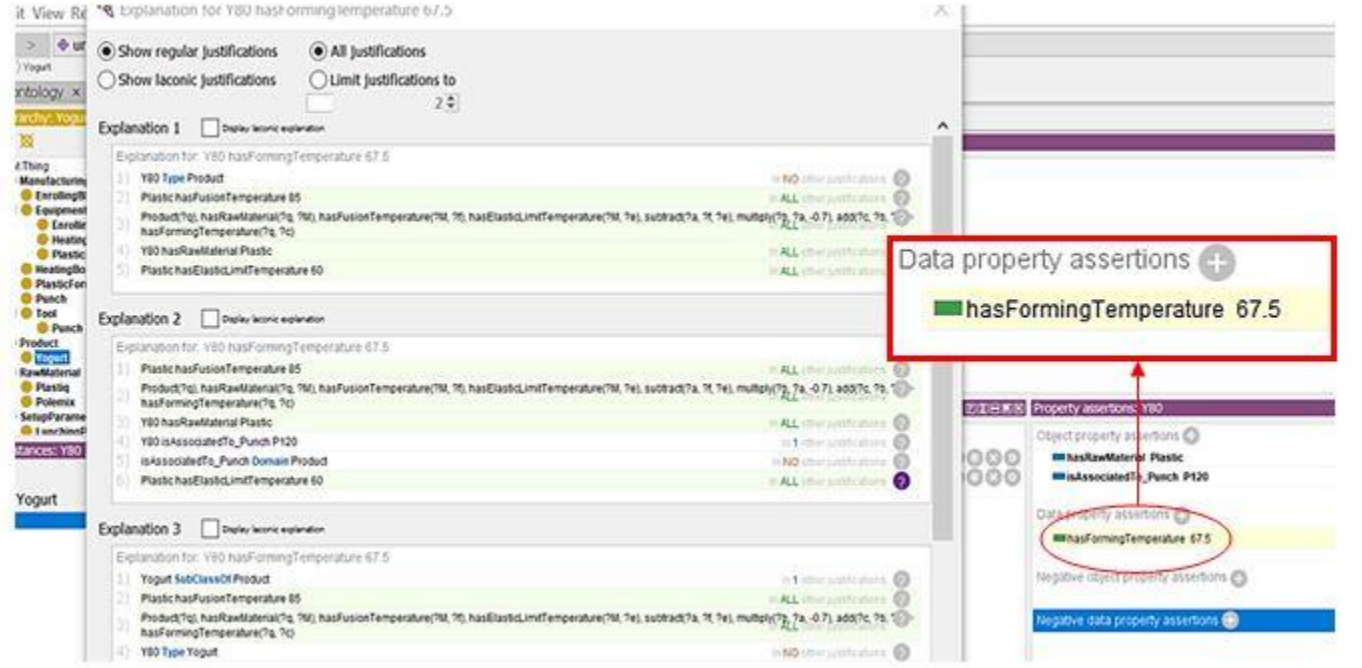


Fig. 12. The Heating Temperature Calculation Rule Encoded in Protégé 5.0, its Results and their Explanation.

All these rules, their results and explanations are shown in Fig. 13. All these rules, their results and explanations are shown in Fig. 13. For instance, the ontology concluded that the dimensional constraint Cd31 of the punch is 10.5 mm.

For the remaining block, i.e. the welding block, the welding parameters are kept the same considering that the SI =100% (the same manufacturing material).

From the results obtained, it is clear that the SPPDT approach has given very satisfying results and therefore it can be validated.

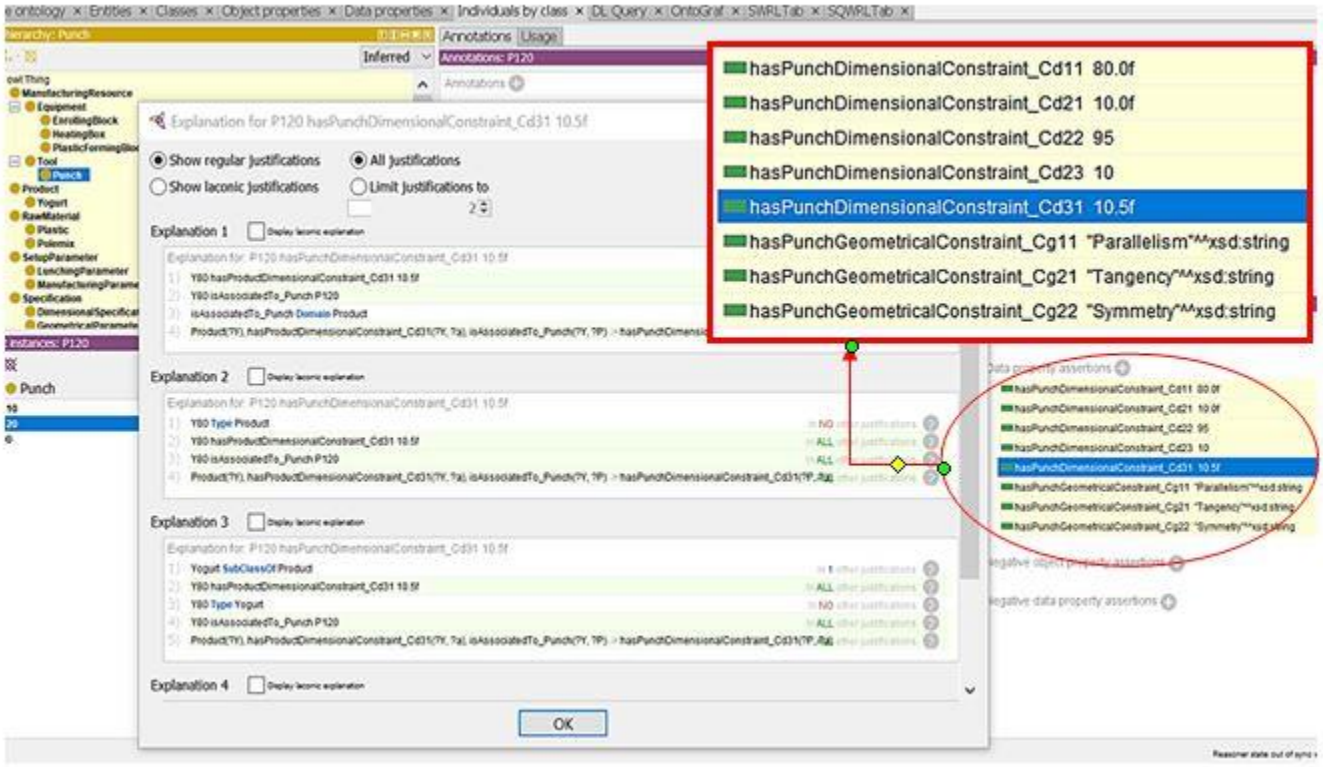


Fig. 13. The Plastic Forming Block Rules Encoded in Protégé 5.0, their Results and Explanations.

VIII. Conclusion and Perspectives

In this paper, a new automated artificial intelligence system is developed to support decision making for selection of production parameters. Its originality lies in the integration, for the first time, of three different artificial intelligence tools, namely: digital twin, ontologies and case-based reasoning. The integration of these tools in a flexible hybrid system allows benefiting from the different advantages of each of them. In fact, the digital twin allowed us to simulate the production processes and their parameters in real time, as well as to validate the production parameters on the virtual production processes before their physical implementation. On the other hand, the use of ontologies allowed us to ensure the interoperability between the different elements of the cyber-physical model of the digital twin, to ensure the expressiveness of the treated information and to preserve their semantics. In addition to that, ontologies and together with CBR ensured reasoning and decision making for the selection of optimal production parameters. Differently from previous works which always present connection limits between the physical and the virtual DT, we have dealt with a double problematic, that of reasoning for the optimal choice of the production parameters, and that of interoperability via the assurance of the physical-virtual connection.

The originality of this work also lies in the efficiency of our ontology to be adapted to several domains, through the formalization of business rules.

To illustrate these advantages and the effectiveness of the developed SPPDT system, an industrial case study is presented at the end of this document. Indeed, the development of this study allowed us to select the optimal production parameters of the studied process in an automated way.

As perspectives, it is suggested to enrich the developed SPPDT system, in particular its DTM-Onto, by integrating other aspects such as the degradation of the equipment, the external factors influencing the production, the safety of the equipment and the personnel, etc. Another perspective is to process and automate the selection of production parameters using other artificial intelligence tools.

References

[1] G. Mezzour, S. Benhadou, and H. Medromi, “Digital Twins Development Architectures and Deployment Technologies: Moroccan use Case”, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 2, 2020.

[2] L. D. Xu, E. L. Xu, and L. Li, “Industry 4.0: state of the art and future trends”, International Journal of Production Research, vol. 56, no. 8, p. 2941-2962, Apr. 2018.

[3] Z. Cunbo, J. Liu, and H. Xiong, “Digital twin-based smart production management and control framework for the complex product assembly shop-floor”, The international journal of advanced manufacturing technology, vol. 96, no. 1-4, p.1149-1163, 2018.

[4] M. Cuc, "Improving the decision-making process by modeling digital twins in a big data environment.", Management & Marketing Journal, vol.19, no.1, 2021.

[5] B. Hamrouni, A. Bourouis, A. Korichi et al., “Explainable Ontology-Based Intelligent Decision Support System for Business Model Design and Sustainability”, Sustainability, vol. 13, no. 17, p. 9819, 2021.

[6] A. Martin, S. Emmenegger, K. Hinkelmann & B. Thönssen, “A viewpoint-based case-based reasoning approach utilising an enterprise architecture ontology for experience management”, Enterprise Information Systems, 11(4), p. 551-575, 2017.

[7] T. R. Gruber, “Towards Principles for the Design of Ontologies Used for Knowledge Sharing in Formal Ontology in Conceptual Analysis and Knowledge Representation”, Kluwer Academic Publishers, 1993.

[8] R. Gámez Díaz, “Digital Twin Coaching for Edge Computing Using Deep Learning Based 2D Pose Estimation”, Thèse de doctorat. Université d'Ottawa/University of Ottawa, 2021.

[9] M Grieves, “Virtually perfect: Driving Innovative and Lean Products Through Product Lifecycle management”, Space Coast Press, Cocoa Beach, Fl. USA, 2011.

[10] M Grieves, J Vickers, “Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems”, Transdisciplinary Perspectives on Complex Systems, Springer International Publishing, Berlin, Germany, 2017.

[11] F. Xiang, Z. Zhang, Y. Zuo et al., “Digital twin driven green material optimal-selection towards sustainable manufacturing”. Procedia Cirp, vol. 81, p. 1290-1294, 2019.

[12] F. Tao, F. Sui, A. Liu et al., “Digital twin-driven product design framework”, International Journal of Production Research, vol. 57, no. 12, p. 3935-3953, 2019.

[13] F. Tao, M. Zhang, Y. Liu et al., “Digital twin driven prognostics and health management for complex equipment”. Cirp Annals, vol. 67, no 1, p. 169-172, 2018.

[14] International Standard Organization, «ISO-14258-1998: Systèmes d’automatisation industrielle, concepts et règles pour modèles d'entreprises» 1998.

[15] V. Fortineau, “Contribution à une modélisation ontologique des informations tout au long du cycle de vie du produit”, Chemical and Process Engineering, Ecole nationale supérieure d'arts et métiers - ENSAM., France, 2013.

[16] G. Declerck, A. Baneyx, X. Aimé et al., “ Les ontologies fondationnelles peuvent-elles débabeliser le web? ”, Rev. d'Intelligence Artif., vol. 28, no 2-3, p. 191-216, 2014.

[17] M. Garetti, L. Fumagalli and E. Negri, “Role of ontologies for CPS implementation in manufacturing”, Management and Production Engineering Review, vol. 6, no. 4, p. 26-32, 2015.

[18] V. Fortineau, T. Paviot and S. Lamouri, “Improving the interoperability of industrial information systems with description logic-based models - the state of the art”, Computers in industry 64, p. 363-375, 2013.

[19] A. Abadi, H. Ben-azza et S. Sekkat, “An ontology-based support for knowledge modeling and Decision-Making in Collaborative Product Design”, International Journal of Applied Engineering Research, vol. 12, no 16, p. 5739-5759, 2017.

[20] A. Matsokis et D. Kiritsis, “An ontology-based approach for product lifecycle management”, Computers in industry, 61, p. 787–797, 2010.

[21] F. Naya, M. Contero, J. Dorribo Camba et al., “On the role of geometric constraints to support design intent communication and model reusability”, 2020.

[22] M. D. De Azevedo Jacyntho, et D. Morais, “Ontology-based decision-making”, In : Web Semantics. Academic Press, p. 195-209, 2021.

[23] M. Horridge,S. Jupp, G. Moulton, A. Rector, R. Stevens and C. Wroe, “A practical guide to building owl ontologies using protégé 4 and co-ode tools edition1”, vol. 2, The university of Manchester, pp. 107, 2009.

Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context

Orieb Abu Alghanam, Sumaya N. Al-Khatib, Mohammad O. Hiari
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—Nowadays e-commerce environment plays an important role to exchange commodity knowledge between consumers commonly with others. Accurately predicting customer purchase patterns in the e-commerce market is one of the critical applications of data mining. In order to achieve high profit in e-commerce, the relationship between customer and merchandise are very important. Moreover, many e-commerce websites increase rapidly and instantly and competition has become just a mouse-click away. That is why the importance of staying in the business, and improving the profit needs to accurately predict purchase behavior and target their customers with personalized services according to their preferences. In this paper, a data mining model has been proposed to enhance the accuracy of predicting and to find association rules for frequent item sets. Also, K-means clustering algorithm has been used to reduce the size of the dataset in order to enhance the runtime for the proposed model. The proposed model has used four different classifiers which are C4.5, J48, CS-MC4, and MLR. Also, Apriori algorithm to provide recommendations for items based on previous purchases. The proposed model has been tested on Northwind trader's dataset and the results archives accuracy equal 95.2% when the number of clusters were 8.**

***Keywords*—Apriori PT algorithm; C4.5; CS-MC4; Data mining; decision tree; e-commerce; K-means**

I. INTRODUCTION

The technique of examining data from a different category is known as data mining [1]. This data contains important information, also in data mining additional knowledge will be extracted. Also, it is a helpful strategy for extracting and detecting patterns in huge data sets that incorporate methods from machine learning, statistics, and database system [2, 3].

Nowadays corporate organization is attempting to adopt a digital marketing strategy and competitive markets in order to gain worldwide commercial benefits. on the other hand, to get such competitive advantages, e-commerce businesses must first comprehend their customers' sentiments, thoughts, and seasons in relation to their products and services [4].

Competitive economy and customer repurchasing behavior are critical to a company's existence. Deeper marketing tactics and managerial decisions can be made with a better grasp of customers and their preferences [5]. A typical online retail store has thousands of transactions in its database, and it serves hundreds, if not thousands, of customers per day. Manipulation and processing of this data in various ways to provide a model with increased prediction accuracy allow for the extraction of novel knowledge that aids in one-to-one marketing, personalization, increased sales, and customer retention [6]. Network marketing has become a significant marketing technique, and as internet technology has advanced, many companies have built online stores to give customers purchasing materials. Because of the numerous benefits of e-commerce, the number of people who engage in online trade, as well as the volume of transactions, has significantly expanded [7].

The difficulty in data mining applications is identifying valid, relevant, and intelligible information from raw and sparse data by mining frequent patterns for knowledge discovery [8]. One of the most important applications of data mining in the e-commerce sector is accurately anticipating client purchase habits because the number of e-commerce websites (both customer and merchandise) grows swiftly and instantly, and competition is only a mouse click away. To stay in business, providers must be able to reliably forecast customer buying behavior and target them with customized services based on their preferences.

Machine learning (ML) techniques are one of the most techniques that are used as data mining techniques. Also, using ML to develop the learner model based on previous experiences and get new knowledge when the size of data becomes huge. ML has been used in many fields such as security [9, 10] medical field, e-commerce field and others.

e-Commerce data is referred to as "Big Data" therefore dealing with this data to extract the knowledge is considered a challenge [11]. In addition to size, using analytical approaches and solutions to extract patterns in hidden relationships in order to make better decisions and get new knowledge makes it more complicated. Furthermore, choosing suitable algorithm to get the best pattern and extract the knowledge to improve the performance is also not easy.

The data mining applications have problems in the mining of recurrent patterns seeking knowledge discovery in order to identify valid, useful and understandable information out of raw and sparse data.

Applying a data mining model to enhance the accuracy of prediction in the context of e-commerce and dealing with big data to extract the knowledge at a reasonable time is an important task. Also, presenting suggestions for associated item set using a prior PT algorithm to help the customer is a desirable task.

The major contributions in this paper are as follows:

1) Applying data mining algorithms in such a way to provide a model that predicts customers' next purchase and recommends it to them.

2) Providing a comparison between different decision tree classification algorithms to choose the best classification algorithm for a product recommendation system based on a set of considered parameters.

3) Clustering the data to enhance the runtime and using Apriori PT association rule algorithm to extract the set of items.

The rest of the paper is organized as follows: Section 2 presents related works and Section 3 presents data collection while Section 4 gives the suggested system framework and major contribution. Moreover, the experiment results are shown in Section 5, and the conclusion is presented in Section 6.

II. Related Work

The rise in popularity of social media has ushered in a new era for e-commerce, transforming online shopping. Several studies have been proposed to enhance the performance for the prediction in e-commerce [12]. Also, some of them used to predict the customer opinion based on the comments [4]. This section presents different classification algorithm that has been used in the literature for data miming or for classifying.

The related data mining algorithm has been presented for e-commerce but from different perspectives. On the other hand, the proposed approaches have differed from the contribution of this paper such as the objectives and the datasets and the way that the proposed model is designed. This section presents the data mining algorithms and how it has been used in the context of e-commerce.

A. K-means

The k-means algorithm is a data mining technique that splits entities into K groups based on attributes or features, where K is a positive integer number [13]. In order to group data, the sum of squares of distances between data and the respective cluster centroid is minimized. K-mean clustering is used to organize data into categories. Fig. 1 shows the K-means algorithm when the number of clusters has been selected to be 5.

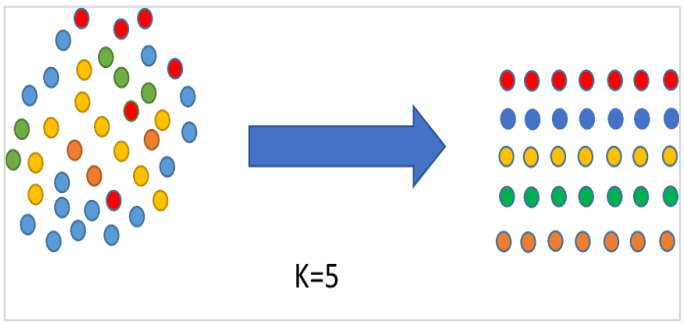


Fig. 1. Dataset Attributes and Types.

Anitha and Patil applied the Recency, Frequency and Monetary model (RFM) and deploy the principles of dataset segmentation using the K-Means Algorithm. This model objective is to employ business intelligence (BI) in recognizing potential customers by providing timely data that is relevant to the retail industry's business units. The used data was based on systematic research and scientific applications in the analysis of sales history and consumer purchasing behavior. The data, carefully selected and organized as a result of this scientific research, not only increase business sales and profits but also provides intelligent insights for predicting consumer purchasing behavior and related patterns. They also used the KMeans clustering algorithm Silhouette Analysis to evaluate the clusters by varying the number of clusters. Based on the Silhouette Score, they concluded that they could analyze the up-to-dateness of the sale, the frequency of the sale, and the money of the sale and find the best solution. [14].

Mulyawan and et al analyzed the behavior of customer shopping by made web shopping. Analyzing and comprehending clients purchasing patterns might assist web shopping in determining what they are seeking. Based on a "transaction" data set, they used Frequency and monetary (FM) analysis. They then divided the clients into groups based on how frequently they bought, how much they bought, and how much the acquired item was worth. They also clustered customers based on their transactions using the K-means method. The study indicates that the K-means algorithm suggestion products were successfully achieved and shown on the customer page [15].

B. Classification by Decision Tree Induction

Databases have plenty of interesting hidden information that can be intelligently used for decision-making. Decision trees are known as classification trees and they are used in machine learning due to their ability to handle both discrete and continuous data in big databases and their easy implementation.

Redouan ABAKOUY and et al employed a learning model for predicting the "clicks" and "conversions" of targeted marketing emails. They compared algorithms of regression and classification for predicting whether an email sent will be opened, clicked, or converted by the intended recipient or not. The features gathered from the emails and client profiles were used to create the model. They compared categorization approaches for predicting whether an email sent to a possible recipient will be opened or not. They are the SVMs classifier and the C4.5 Decision Tree classifier. In all the cases, the Decision Tree classifier results outperform the SVM. [16].

For e-commerce logistics businesses to manage enormous client bases and develop long-term and profitable connections, Luk and et al presented an intelligent customer identification model (ICIM). This ICIM comprises a historical view and analysis of all existing or potential consumers. That model aided in the accurate identification of actual consumer needs, as well as the classification of new clients in the future in the shortest period possible. The ICIM combines the k-means clustering technique and the C4.5 classification algorithm to extract important hidden knowledge from both continuous and discrete variables [17].

1) C4.5 Decision tree: C4.5 is an improved version of the greedy, top-down, recursive, divide-and-conquer ID3 algorithm; the improvement in the algorithm included its ability to handle continuous variables, prune the tree after being created and its ability to deal with missing values. C4.5 rules are then constructed by greedily prune conditions from each rule if this decreases its estimated error.

2) Improved J48 decision tree classification algorithm: The J48 algorithm is a well-known machine learning algorithm that is based on the J.R. Quilan C4.5 algorithm [18]. In this paper, the algorithm is evaluated against C4.5 for verification purposes. With this technique, a tree is built to model the categorization process using this technique. Once the tree has been constructed, it is applied to each tuple in the database, yielding categorization for that tuple.

3) CS-MC4 Decision tree algorithm: The main goal of decision tree induction algorithms is to increase accuracy while minimizing costs. The m-estimate smoothed probability estimation process, which is a generalization of the Laplace estimate [19], is used in the cost sensitive decision tree algorithm. This approach decreases the expected loss by detecting the best prediction within leaves using a misclassification cost matrix.

Table I presents a comparison between different approaches that have been proposed for e-commerce. The comparisons have been done in terms of the algorithm that is used, the datasets and the experiment results. On the other hand, in this paper different data mining model has been proposed that aims to enhance the prediction in addition to apply the prior algorithm to generate a rule for association items.

TABLE I. COMPARISON BETWEEN DIFFERENT DATA MINING APPROACHES

Reference	Dataset	Proposed approach	Results
[20]	Amazon DVD musical product:- .Net Crawler, 9555 reviews	Hybrid approaches	Precision: - 0.89, Recall: - 0.84, F-Measure: - 0.86
[21]	Review of cellphone & accessories:- 21600 reviews [22]	Linear support vector machine	Accuracy: - 93.52%
[11]	UCI Machine Learning Repository.	Decision Tree	Accuracy: - 95%
[23]	Data in [24]	C4.5	Accuracy: - 86.59%
		Random forest	Accuracy: - 86.78

C. Apriori PT Association Rule Algorithm

Yuanzhu and et al present a study that implements the Apriori algorithm and C5.0 which are considered as association rules; also, decision tree techniques for data mining [25]. It has been used to help managers or decision maker people to extract knowledge ‘from’ and 'about' customers in order to determine their preferences, allowing enterprises to develop the correct goods and achieve a competitive advantage.

The findings show that the knowledge-based approach is effective, and the returned knowledge is represented as a set of rules that can be used to identify relevant patterns for both new product development and marketing tactics.

The original Apriori algorithm was proposed by Agarwal and Srikant in 1994 [26]. Apriori is constructed to operate on transactional databases; the algorithm determines item sets in the database that are subsets of at least one transaction. Apriori PT is an enhancement of this algorithm that works well with big data by:

Step 1 - find all frequent elements that have support more than the minimum support needed. Step 2 - the set of frequent elements to build association rules with a high enough level of confidence.

Pruning -using the fact that any subset of a frequent item set should be frequent.

III. DATA COLLECTION

The real lifetime large transactional data set from Kaggle to test the purposed model. The northwind.mdb sample transaction database has been used to test the robustness of the proposed model. The dataset passed through preprocessing phases to meet the requirements for each selected algorithm such as data type conversion from continuous to discrete was also handled for prediction purposes.

Northwind originally consisted of many tables with relations between them, each table consists of many details. In this paper, processing and understanding the dataset have been done. Furthermore, 2155 product sales on 8 product categories of 77 different item types for 91 different customers have been taken into consideration. Also, the demographic variable such models, gender and customer job have been taken. Table II presents the dataset details while Fig. 2 contains further dataset details.

TABLE II. CHARACTERISTICS NORTHWIND DATABASES

Attribute	Category	Information
Customer ID	Discrete	89 values
Gender	Discrete	2 values
Customer Job	Discrete	12 values
Order ID	Continue	-
Category Name	Discrete	8 values
Product Name	Discrete	77 values
Unit price	Continue	-
Quantity	Continue	-
Discount	Continue	-
Extended Price	Continue	-

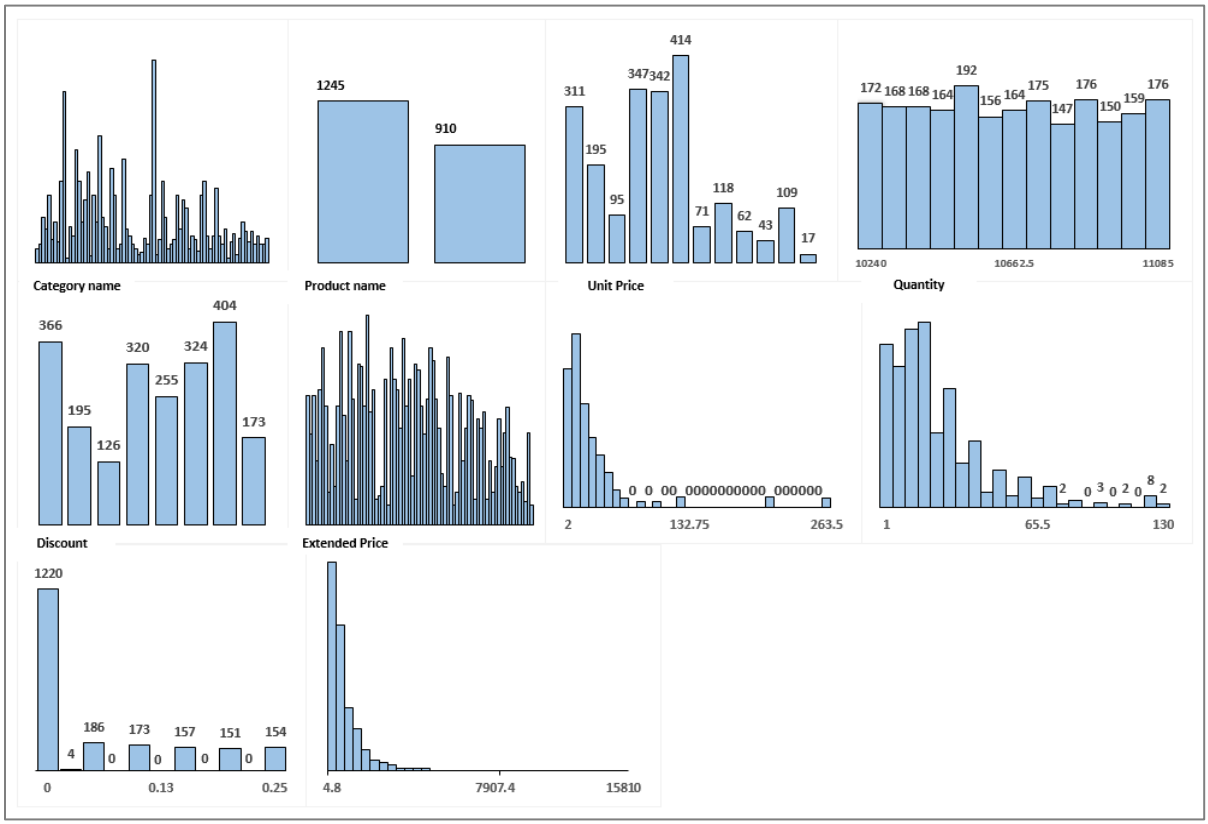


Fig. 2. Dataset Attributes and Types.

IV. System Framework

Fig. 3 presents the proposed model which went through a group of phases. In the first phase a normalization and duplicate removal have been done, then k-means clustering was performed in the second phase. In the next phase, the data was split into training and testing. Moreover, the modeling stage has been built based on four algorithms which are C4.5, J48, CS-MC4, and MLR. Also, a prior PT algorithm has been applied to get the association rules.

The proposed system firstly starts by deciding how many K clusters need to split the dataset. The centroid or center of these clusters has been randomly chosen to start the calculation for the whole data. Moreover, to divide the dataset into clusters this will be done based on the distance between each object and the centroid to categorize the objects based on the minimum distance (closest centroid). Table III presents the characteristics of the K-means clustering algorithm and the number of clusters that are generated in the proposed model.

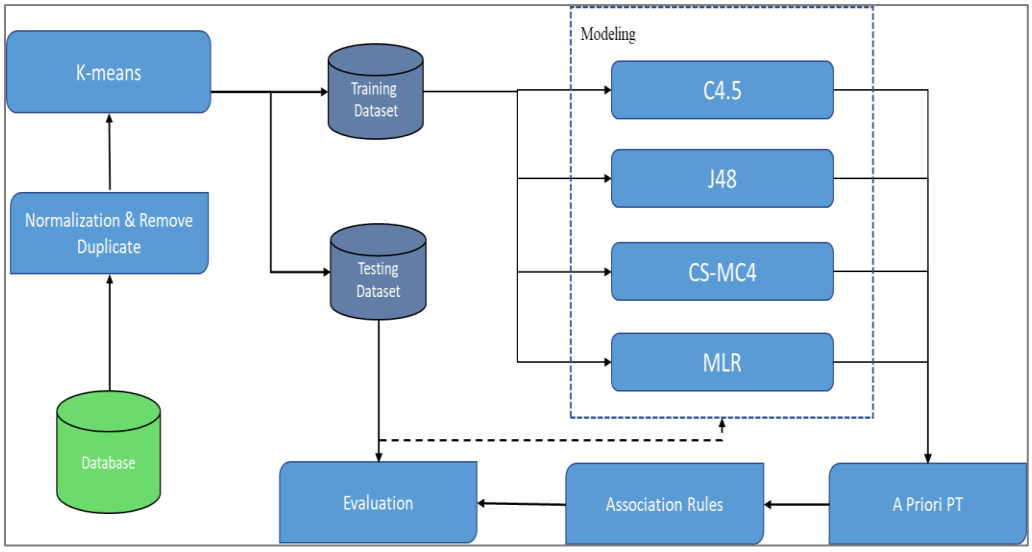


Fig. 3. The Proposed Data Mining Model.

TABLE III. K-MEANS CLUSTERING PARAMETERS

Name	Description	Value
Distance function	The distance function to use for instances comparison.	Fixed - Euclidean distance
Number of clusters	The number of clusters to be created.	Dynamic – starting with 2 clusters and finished at 12 K= 2, 8, 10,12
Max Iterations	Set the maximum number of iterations.	Fixed - 500
Seed	The random number seed to be used.	Fixed - 10

In this paper, the dataset has been clustered into different clusters which are 2,8,10,12. Furthermore, the maximum iteration that is used is 500 while the fixed Euclidean distance is used for instances comparison between the records. Also, the random number called seed that is used is 10.

The Euclidean distance or the Manhattan distance is used to cluster data when utilizing the K means technique as shown in equation 1. If the Manhattan distance is employed, the component-wise median rather than the mean is used to calculate the centroids [27].

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

Classification's fundamental goal is to accurately anticipate the target class for each record. Its training procedure seeks to uncover correlations between predictor and target variables.

Classification algorithms [28, 29] differ in the strategies employed to identify these associations, which are further summarized in a model then applied to a record (test data) where the class label is unknown.

The modeling stage is built based on four algorithms which are C4.5, J48, CS-MC4, and MLR. Each algorithm has been applied on the whole clustered dataset, then substituting error rates for each. Moreover, unbiased error rate estimation ‘10 folds cross-validation was used to evaluate each learning algorithm. Table IV presents the parameters that have been used by C4.5 algorithm and the splitting ratio for the dataset.

In the proposed model the weighted total of the error estimates for all of the subtree's leaves has been used to get the error estimate. The upper bound of the error estimate for a node is derived as shown in equation 2, where f represents the error on the training data and N is the number of instances covered by the leaf.

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right) \quad (2)$$

TABLE IV. C4.5 DECISION TREE PARAMETERS

Minimum Size of Leaves	5
Confidence level - lower values incur heavier pruning	25%
Cross Validation	10 folds
Train-Test	80% (train) 20% (test)

The parameters for J48 algorithm that have been used in the proposed model are presented in Table V. more details about these parameters in [30]. The default Number of Folds has been used which is 3 and the seed was 1.

TABLE V. J48 TREE PARAMETERS

Collapse Tree	yes
Confidence Factor	25%
Min. No. of Objects	2
Number of Folds	3
Seed	1
Use MDL Correction	True
Unpruned	False
Cross Validation	10 folds
Subtree Raising	True

Table VI shows the parameters applied to implement CSMT4 algorithm.

TABLE VI. CS-MC4 CLASSIFICATION TREE SUPERVISED PARAMETRS

Minimum Size of Leaves	5
Lambda	3
Cross Validation	10 folds

In the next step, Apriori PT Christian Borgelt’s as shown in Fig. 4 was applied, which is a highly effective association rule generator, it can handle large datasets quickly. Further processing has been done before using Apriori algorithm. The processed data consisted of 830 transactions and 77 attributes. The "item types” was set to ‘yes' for each item purchased by each transaction and 'no' otherwise. Also, the main support was set at 0.1 (10%), the min confidence min as 85%, the max cardinal of the item was set as 4 (Max Card Item sets).

I_k: frequent item set of size K

Li= {frequent item}

For (K=1; I_k !=0 ; K++) do begin

C_k +1= candidates generated

From I_k;

do

Increment the count of all

Candidates in C_k+1

I_k +1= Candidates in C_k +1 with

Min_support

End

Return: K, I_k

Fig. 4. Input Data for Apriori PT Algorithm and Algorithm Pseudocode.

The association rules were generated based on the proposed model. Here is a sample set of generating rules which are related to the attribute number. If attribute 39 and attribute 77 are combined, then the attribute for 46 should represent the product reality-lifetime and unique ids this means if a customer purchased item number 39 and item number 77 then item 46 most probably will be bought. Thus, the proposed model recommends this item to that customer. Finally, to test the usability of the proposed model we applied the model to the real-life time usability of big Dataset and the model showed high robustness.

One of the strong features of C4.5 algorithm is its ability to handle discrete and continuous data types, this feature was used and the algorithm was implemented on clustered data, with 96.9 % accuracy.

V. Experiments and Results

This section presents the experimental results for the proposed model. Moreover, each of the above-mentioned decision trees is executed for each learning method on the whole dataset. After that, we substituted error rates for each, then we use unbiased error rate estimation '10 folds cross-validation' to evaluate each learning algorithm.

A. The Performance Measurement

In this paper, the accuracy and the error rate have been used to measure the performance of the proposed model and measure the performance of each classifier.

1) Accuracy: It is referring to the proportion of valid predictions (including true positives and true negatives) among the total number of cases analyzed is the accuracy [31]. Classified by the classifier as shown in equation 3:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2) Error-Rate: It's also known as the Misclassification rate, and it's calculated as 1-Acc (M), where Acc (M) represents M's accuracy, as given in equation. 4:

$$Error\ Rate = 1 - \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Table VII shows the experiment results for all classifiers in terms of the accuracy and error rate for each classifier on different clusters size. The experiment results have been done on the dataset based on clustering it into a different cluster. K-means algorithm has been used to reduce the size of the dataset and enhance the speed and performance of the classifiers.it can be noticed that when the size of the cluster was 10 the best accuracy has been reached compared with other clusters size.

Moreover, C4.5 outperforms the other classifiers in terms of accuracy in all clusters size. On the other hand, CS-MC4 has reached the lowest accuracy compared with the other classifiers in all clusters size.

The results turn out that using 10 clusters gave petter results for the C4.5 algorithm that is reach 96.9% accuracy while when 8 clusters have been used MLR reach 89%. C4.5 can handle discrete and continuous data types; it is used this strong point feature and implemented the algorithm on clustered data.

When j48 induction tree algorithm has been applied to '10' k-means clustering dataset the prediction accuracy was 93.8%. The testing has been performed in the 10-fold cross validation. After that, the results are then used to generate decision rules.

TABLE VII. The Comparisons between Different Algorithms for Different Clusters Numbers

K-means																
# of clusters	**K=2**				**K=8**				**K=10**				**K=12**			
Algorithm	**C4.5**	**J48**	**CS-MC4**	**MLR**	**C4.5**	**J48**	**CS-MC4**	**MLR**	**C4.5**	**J48**	**CS-MC4**	**MLR**	**C4.5**	**J48**	**CS-MC4**	**MLR**
Accuracy%	86.5	82.4	62.7	70.6	95.2	92.2	81.7	89.3	69.9	93.8	83.4	91	93.1	90	79	87.1
Error Rate %	13.5	17.6	37.3	29.4	5	7.8	18.3	10.7	3.1	6.2	16.6	9	6.9	10	21	12.9

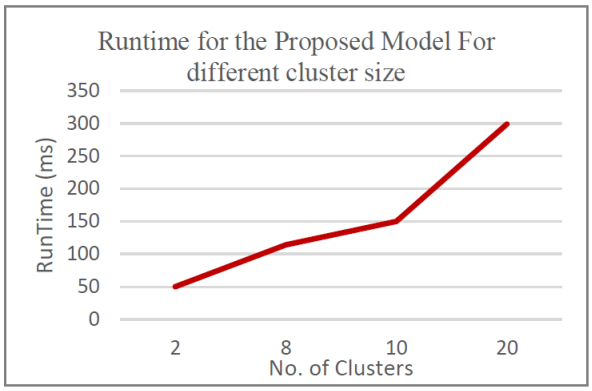


Fig. 5. The Runtime for the Proposed Model with different Clusters.

Fig. 5 shows the results for the runtime that is needed when the different size of clusters has been used. It can be noticed that when the number of clusters increases the runtime is increasing. It can be noticed that when the data has clusters into any two clusters the runtime that has been taken was 50msWhile when the number of clusters becomes 12 the runtime reached approximately 7 times greater than when the data was 2 clusters.

Table VIII illustrated the size of the decision tree in terms of the number of leaves and number of nodes that are generated by C4.5 algorithm, J48 algorithm and CS-MC4 algorithm for the dataset in terms of the number of nodes and the number of leaves.

TABLE VIII. NUMBER OF NODES AND LEAVES FOR EACH CLASSIFIER

Algorithm	C4.5	J48	CS-MC4
No. of nodes	35	57	83
No. of Leaves	18	23	45

Fig. 6, Fig. 7, Fig. 8 and Fig. 9 represent the accuracy for different algorithms based on different cluster sizes. The clusters that have been selected were two, eight, ten and twelve respectively. It can be noticed that when the cluster size was 10 all classifiers reach a better accuracy compared with other cluster sizes. Also, the results for C4.5 algorithm outperforms the other algorithms in all cluster size.

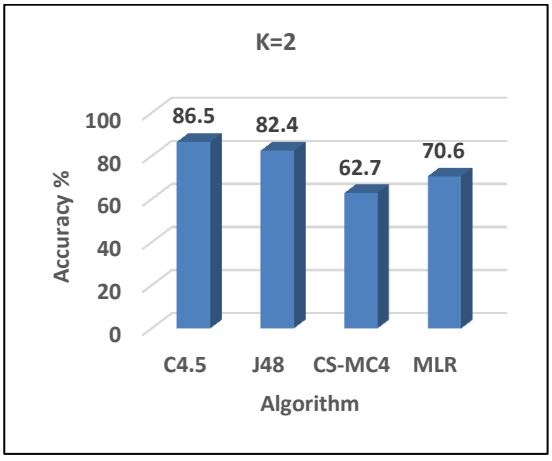


Fig. 6. The Accuracy for different Classifiers when k=2.

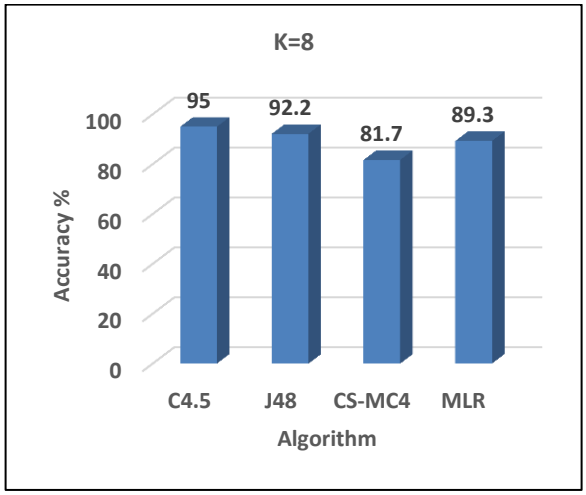


Fig. 7. The Accuracy for different Classifiers when k=8.

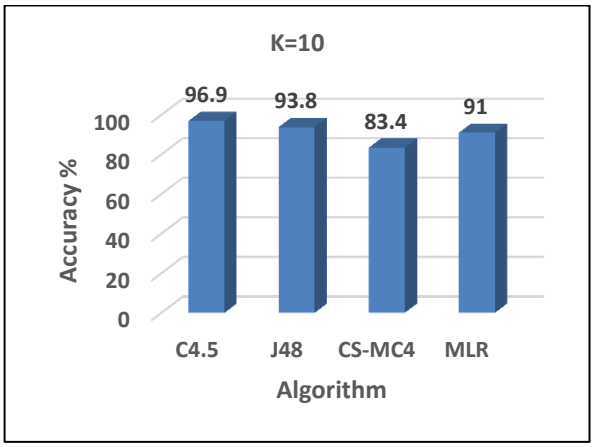


Fig. 8. The Accuracy for different Classifiers when k=10.

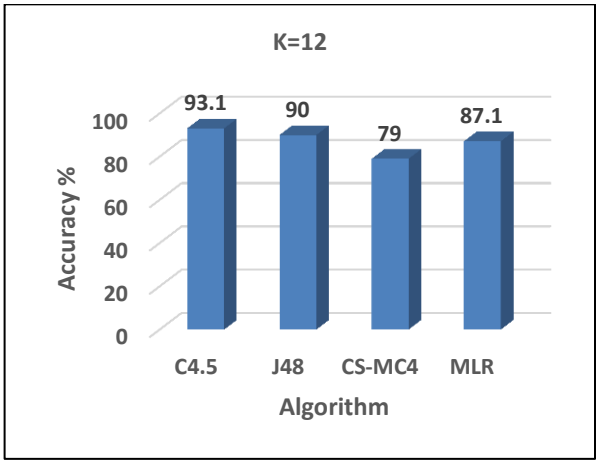


Fig. 9. The Accuracy for different Classifiers when k=12.

VI. CONCLUSION

In conclusion, this paper achieved higher accuracy for predicting in the proposed data mining model. Also, it gives an indicator for the suitable size of the clusters that should be selected for northwind.mdb. Moreover, the proposed model suggests the most association items that are related to each

other. It amid to understand the purchase behavior to predict customer next purchase based on a set of selected parameters when Apriori PT algorithm has been used. On the other hand, the proposed model aimed to enhance the prediction for a huge database. The experimental results show that J48 and C4.5 algorithms produce high accuracy measurements compared with other algorithms.

In this paper, Apriori PT is applied for a fast and powerful association rules generation in e-commerce customer purchasing field. Moreover, data clustering has provided a good performance, such as the run time of the proposed model or the accuracy. Clustering the dataset does not affect the value of the data. Finally, the proposed model achieved 95.2% accuracy when the number of clusters was assigned to be 8 for C4.5 algorithm. On the other hand, the CS-MC4 algorithm achieved the lowest accuracy when the number of clusters was 2 it reached 62.7%.

REFERENCES

[1] S. F. Abdullah, A. F. N. A. Rahman, Z. A. Abas, and W. H. M. Saad, "Fingerprint gender classification using univariate decision tree (j48)," Network (MLPNN), vol. 96, no. 95.27, pp. 95-95, 2016.

[2] T. Reutterer, M. Thomas, and N. Schröder, "Leveraging purchase regularity for predicting customer behavior the easy way," International Journal of Research in Marketing, vol. 38, no. 1, pp. 194-215, 2021.

[3] R. Heldt, C. S. Schmitt, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," Journal of Business Research, vol. 127, pp. 444-453, 2021.

[4] A. Moazzam, Y. Farwa, H. Mushtaq, A. Sarwar, A. Idrees, S. Tabassum, BaburHayyat, and K. Ur Rehman, "Customer Opinion Mining by Comments Classification using Machine Learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 5, pp. 385-393, 2021.

[5] K. Maheswari, and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), pp. 1-5, IEEE, 2017.

[6] X. Dou, "Online purchase behavior prediction and analysis using ensemble learning," in 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pp. 532-536, IEEE, 2020.

[7] R. Heldt, C. S. Silveira, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," Journal of Business Research, vol. 127, pp. 444-453, 2021.

[8] A. Dogan, and D. Birant, "Machine learning and data mining in manufacturing," Expert Systems with Applications, vol. 166, pp. 114060, 2021.

[9] O. AbuAlghanam, L. Albdour, and O. Adwan, "Multimodal Biometric Fusion Online Handwritten Signature Verification using Neural Network and Support Vector Machine," International Journal of Innovative Computing, Information and Control, vol. 7, no. 5, pp. 1691-1703, 2021.

[10] S. N. Mohanty, E. L. Lydia, M. Elhoseny, M. M. G. Al Otaibi, and K. Shankar, "Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks," Physical Communication, vol. 40, pp. 101097, 2020.

[11] E. F. Zineb, N. RAFALIA, and J. ABOUCHABAKA, "An Intelligent Approach for Data Analysis and Decision Making in Big Data: A Case Study on E-commerce Industry," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 7, pp. 723-736, 2021.

[12] Y. Fu, M. Yang, and D. Han, "Interactive Marketing E-Commerce Recommendation System Driven by Big Data Technology," Scientific Programming, vol. 2021, 2021.

[13] T. Mitchell, Machine Learning. McGraw hill Burr Ridge, 1997.

[14] P. Anitha, and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorith," Journal of Kings Saud University-Computer and Information Sciences, 2019.

[15] B. Mulyawan, M. V. Christanti, and R. Wenas, "Recommendation Product Based on Customer Categorization with K-Means Clustering Method," IOP Conference Series: Materials Science and Engineering, vol. 508, no. 1, pp. 012123, 2019.

[16] R. Abakouy, E. M. En-naimi, A. E. Haddadi, and E. Lotfi, "Data-driven marketing: how machine learning will improve decision-making for marketers," in proceedings of the 4th international conference on Smart City Applications, pp. 1-5, 2019.

[17] C. C. Luk, K. L. Choy, and H. Y. Lam, "Design of an intelligent customer identification model in e-Commerce logistics industry," MATEC Web of Conferencs, vol. 255, pp. 04003, 2019.

[18] Y. Zhan, K. H. Tan, and B. Huo, "Bridging customer knowledge to innovative product development: a data mining approach," International Journal of Production Research, vol. 57, no. 20, pp. 6335-6350, 2019.

[19] S. M. Karst, R. M. Ziels, R. H. Kirkegaard, E. A. Sørensen, D. McDonald, Q. Zhu, R. Knight, and M. Albertsen, "High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing," Nature methods, vol. 18, no. 2, pp. 165-169, 2021.

[20] U. A. Chauhan, M. T. Afzal, A. Shahid, M. Abdar, M. E. Basiri, and X. Zhou, "A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews," World Wide Web, vol. 23, no. 3, pp. 1811–1829, 2020.

[21] T. U Haque, N. N. Saber, and F. M. Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews," in 2018 IEEE international conference on innovative research and development (ICIRD), pp. 1-6, IEEE, 2018.

[22] R. He, and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one- class collaborative filtering," in proceedings of the 25th international conference on world wide web, pp. 507-517, 2016.

[23] K. Baati, and M. Mohsil, "Real-time prediction of online shoppers' purchasing intention using random forest," in IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 43-51, Springer, Cham, 2020.

[24] C. O. Sakar, S. O. Polat, M. Katircioglu, Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," Neural Computing and Applications, vol. 31, no. 110, pp. 6893-6908, 2019.

[25] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda," in International Workshop on New Frontiers in Mining Complex Patterns, pp. 119-136, Springer, Cham, 2019.

[26] M. H. W. Ho, and H. F. Chung, "Customer engagement, customer equity and repurchase intention in mobile apps," Journal of business research, vol. 121, pp.13-21, 2020.

[27] S. Yang, and G. M. Allenby, "Modeling interdependent consumer preferences," Journal of Marketing Research, vol. 40, no. 3, pp. 282-294, 2003.

[28] J. Qiu, Z. Lin, and Y. Li, "Predicting customer purchase behavior in the e-commerce context," Electronic commerce research, vol. 15, no. 4, pp. 427-452, 2015.

[29] S. Moon, and G. J. Russell, "Predicting product purchase from inferred customer similarity: An autologistic model approach," Management Science, vol. 54, no. 1, pp. 71-82, 2008.

[30] K. Kang, and J. Michalak, "Enhanced version of AdaBoostM1 with J48 Tree learning method," arXiv preprint arXiv:1802.03522, 2018.

[31] N. Kavha, and S. Karthikeyan, "Customer Buying Behavior Analysis: A clustered Closed Frequent Itemsets for Transactional Database," International Journal of Computer Science Issues (IJCSI), vol. 10, no. 3, pp. 113, 2013.

An Effective Analytics and Performance Measurement of Different Machine Learning Algorithms for Predicting Heart Disease

S. M. Hasan Sazzad Iqbal, Nasrin Jahan, Afroja Sultana Moni, Mst. Masuma Khatun
Department of Computer Science and Engineering
Pabna University of Science and Technology
Pabna, Bangladesh

***Abstract*—This Heart disease means any condition that affects to directly heart. Globally, Heart disease is the main reason for death. According to a survey, approximately 17.9 million people died from heart disease in 2019 (representing 32 percent of global deaths). The number of people dying is increasing at an alarming rate every day. So it is necessary to detect and prevent heart disease as soon as possible. Medical experts who work inside the field of coronary heart sickness can predict the rate of coronary heart disorder up to 69%, which is not so useful. Because of the invention of various machine learning techniques, intelligent machines can predict the chance of heart disease up to 84%, which will be helpful to prevent heart disease earlier. In this paper, for picking essential characteristics among all features in the dataset, the univariate feature selection approach was employed. One-of-a-kind machine learning algorithms like K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine were used to assess the performance of these algorithms and forecast which one performs best. These machine learning approaches require less time to predict disease with more precision, resulting in the loss of valued lives all around the world.**

***Keywords*—Machine learning; heart disease prediction; KNN; Naive Bayes; decision tree; random forest; support vector machine**

I. Introduction

Health is considered as a whole state of physical, mental, and social well-being where there's an absence of disease and infirmity. Health may be evolved through doing several activities like a physical workout, adequate sleep, and using utilizing employing through averting unhealthful sports like smoking or immoderate stress. Because of carelessness, health is being affected by various kinds of diseases. As it is known that the human heart is one of the most essential organs in the human body [12]. The average human heart beats 72 beats in step with a minute and pumps about 2000 gallons of blood to each and each part of the human body. But somehow, if the heart is affected by several diseases, then it'll be harmful to the human body, and sometimes it'll cause death also. Nowadays, heart disease is increasing at an alarming rate. Medical professionals can't get an accurate result of heart disease prediction by following a custom. With the assistance of machine learning algorithms, the prediction can be increased and many people can get alert about their disease and can also take preventive actions before it is too late. With the help of machine learning strategies, it's far feasible to collect information from a massive quantity of information and by training the dataset, the machine can predict the result. So it reduces the extra burden on medical professionals. As in the modern world, it can't be imagined daily lives without technology, machine learning has made life easier by predicting and providing proper guidelines about disease. By using machine learning techniques, millions of lives can be saved by predicting disease quickly and providing quicker service to the patients.

A. Problem Statement

From previous research, it came to know that they examined different machine learning techniques. These studies concentrated on a specific impact of machine learning techniques rather than on their optimization. Some researchers experimented with hybrid optimization techniques. The initial stage in this effort is to apply a correlation-based feature selection method. Among all the attributes of the dataset, only the correlated datasets are segmented and this is called the feature selection method. It is a preprocessing method of machine learning which eliminates irrelevant data and increases learning accuracy. To increase classification accuracy, the best subset of features is chosen from all of them. Different machine learning methods are applied to the entire dataset after it has been divided into train and test datasets. After the comparison of different algorithms, identify the algorithm which performs best for predicting heart disease.

II. Related Work

Lots of work has been done in the field of predicting heart disease in previous years. They have attained different levels of accuracy by applying different machine learning techniques. Some of them are given below:

The identity of coronary heart sickness, diabetes with the assistance of neural networks turned into brought by Niti Guru [1]. Experiments had been finished on a sampled dataset of affected person's records. The neural network changed into educated and tested with 13 input capabilities. The supervised set of rules changed into used for the diagnosis of heart sickness. The backpropagation algorithm was used for training data. Whenever any unfamiliar data was inserted, the process identified the unknown data as compared to training data and produced a probability of heart disease.

Another prediction was introduced by M.Sultana, A.Haider, and Mohammad Shorif Uddin [2]. They have illustrated that datasets that are available for heart disease are in the form of the raw datasets and are inconsistent. They extracted the crucial features from the dataset. By using this method, the time complexity and work of the training algorithm were reduced and the accuracy of the proposed model increased. They have worked with Bayes Net and SMO classifiers which are more optimal than NLP and KStar. They have collected datasets from WEKA software and measured performance by running algorithms (Bayes Net and SMO). Then compared the result with predictive accuracy, ROC Curve, and ROC Value.

An optimization of function was performed to acquire better accuracy the usage of Decision Tree by M.A.Jabbar, B.L. Deekshatu, and P. Chandra [3]. It turned into a method for early detection of coronary heart disease.

K.C.Tan, E.J. Teoh proposed a hybrid method of machine learning where Support Vector Machine and Genetic Algorithm were combined [4]. The LIBSVM and WEKA facts mining tools have been used to investigate the result. They've used five datasets for this project. After applying the SVM and Genetic algorithm, they obtained an accuracy of 84.07%.

G. Parthiban and S.K. Srivastava diagnosed coronary heart sickness in diabetic patients using Naive Bayes and SVM algorithms [5]. A record of 500 patients was used to make a prediction. After applying both the algorithms, Naive Bayes provided an accuracy of 74%, and SVM provided an accuracy of 94.60%.

V. Chaurasia and S. Pal experimented on diverse records mining techniques to come across coronary heart disease [6]. The WEKA data mining tools were used right here. They used Naive Bayes, J48, and bagging for this cause. They took a dataset of 76 attributes. Among them, they selected only 11 attributes to make a prediction. Of the three classification algorithms, bagging provided better accuracy to make a prediction.

Fahd Saleh Alotaibi proposed a version comparing among 5 one of a kind machine learning algorithms [7]. In this research, he used Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM algorithms. Among them, the Decision Tree gave the best accuracy.

After reviewing the above papers, the main concept in the back of the proposed gadget is to make a prediction of heart disease primarily based on the given entered facts. For this purpose, we've used KNN, Naive Bayes, Decision Tree, Random Forest, and SVM set of rules primarily based on their accuracy.

III. Methodology

Exploring classification techniques and performing performance analysis, our suggested model predicts heart disease. Our primary goal is to determine whether or not a patient has heart disease. The flow chart depicts the full procedure, Fig. 1.

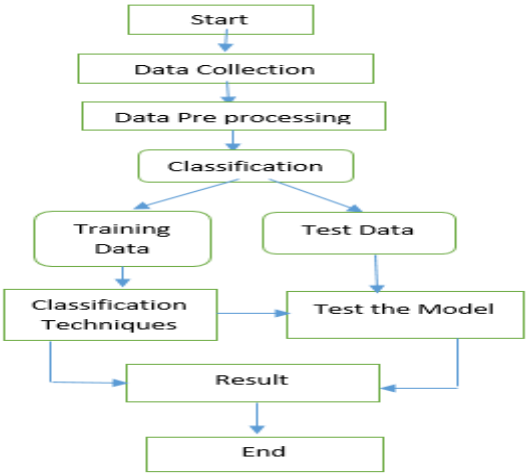


Fig. 1. A Model Predicting Heart Disease.

A. Data Collection and Preprocessing

The dataset we utilized to predict heart disease is a 16-attribute dataset containing 4241 patient records. After gathering data, dimensionality reduction is used. It entails feature extraction and feature selection. The data we've collected has several features or dimensions, but not all of them are necessary or important in terms of the model's output. A huge number of attributes may have an impact on the computational complexity, resulting in a poor outcome.

B. Feature Selection

This approach selects a subset of the original features. The best features are chosen using a univariate feature selection method. The main features are chosen using the chi-square statistics test. In our project, we first standardized our data and then selected the important features. After performing the task, the selected attributes decreased from 16 to 6.

IV. Algorithms and Techniques Used

Because heart disease prediction is a classification or clustering problem, it can be reduced to a single classification with a small number of attributes. It will be easier to identify the correct class as a consequence of this classification challenge, and the result will be more accurate than clustering. We've spoken about the theoretical context that we employed during the experiment in this part. To forecast the best result, we applied five different machine learning methods. Based on their popularity, the KNN, Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine techniques are utilized.

A. K-Nearest Neighbor

K-Nearest Neighbor rule was introduced by Hodges et al. in the year 1951 [8]. He introduced it as a "non-parametric technique for pattern classification". Which is popularly known as K-Nearest Neighbor Algorithm. It is a completely effective and popularly used classification algorithm. The KNN technique may be used for each regression and classification, but it is usually applied for classification tasks. When there's less or no prior knowledge approximately information distribution, it is then used. This set of rules reveals the k-

nearest statistics points inside the training set to the information point for which a goal value is unavailable. Then it assigns the common fee that it has found information factors to the new predicted factor. K-NN is certainly a non-parametric algorithm, which means that it makes no guesses about the underlying records. It's referred to as a lazy learner set of rules because it doesn't learn from the training set at once; rather, it saves the dataset and performs a motion on it when it comes time to categorize it. During the training segment, the KNN algorithm just shops the dataset, and while it accepts new information, it classifies it into a class that's quite similar to the new information [9]. For instance, there are two classes, Category A and Category B, and we get some other new records factor x1. Which of those categories will this information factor suit? A KNN set of rules is needed to solve this sort of trouble. We can readily find out the category or class of a dataset with the assistance of KNN.

B. Naive Bayes

Naive Bayes is an easy classification algorithm that is based totally on Bayes' theorem. It is a collection of classification algorithms and thus it is called a family of algorithms. Naive Bayes assumes independence between the features of the data. It is useful for very large datasets and easy to build. It mainly works depending on probability [10]. It provides calculation posterior probability P(c|x) from P(c), P(x) and P(x|c). The equation that is used to calculate the probability is given below:

$$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$$

Or, it can be represented as:

$$\text{Posterior Probability}=\frac{\text{Likelihood}*\text{class prior probability}}{\text{Predictor Prior Probability}}$$

$$P(c|x)= P(x1|c)\times P(x2|c)\times \ldots\ldots\ldots \times P(xn|c) \times P(c)$$

Here,

- P(c|x) is the posterior probability of the target class.
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor of the given class.
- P(x) is the prior probability of the predictor.

Naive Bayes is easy and fast to estimate the test information set's elegance. It's also proper at multi-class prediction. About to with concerning numerical input variables, it scores well with specific input variables. When the expectancy of independence is met, a Naive Bayes classifier outperforms different fashions.

C. Decision Tree

A Decision Tree is a supervised learning algorithm that is normally used in classification troubles. Here, data is constantly cut up according to a parameter. It is a tree-structured classifier where the decision tree is represented as decision nodes and leaves. The internal nodes of the decision tree denote the features of the dataset, the leaf node denotes the output and each branch denotes decision rules. The decision tree solves the problems by representing them graphically to get all the possible solutions. A decision tree simply continues its splitting process by asking a question whether the answer is positive or negative. Based on its answer, it cut up the trees into subtrees. The basic structure of a decision tree is depicted in the Fig. 2.

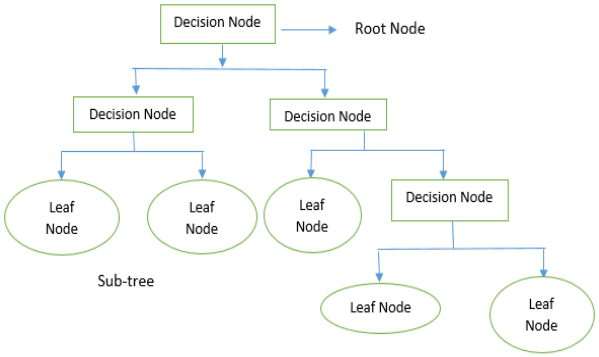


Fig. 2. Decision Tree.

When making any decision, decision trees usually mimic human thinking ability, which is easy to understand. It also uses a tree-like structure for making a decision [11]. The entropy of each characteristic is initially calculated by the decision tree algorithm. It calculates a feature's information and it is known as information gain. Based on information gain, we split the node to make a decision tree. Which characteristic has the highest value is split first, according to the value of information gain. The formula which it follows is:

Information gain= Entropy(S) – [(Weighted avg)*Entropy (each feature)]

Here, entropy is the measurement of impurity in a given attribute. Entropy can be measured as:

Entropy(S) = -P(yes)log 2 P(yes) – P(no)log 2 p(no)

Where,

S= Total no of samples

P(yes) = Probability of yes

P(no) = Probability of no

Decision trees can generate rules that are simple to understand. Classification is achieved via decision trees, which do not require a whole lot of computing. Decision trees indicate which fields are most relevant for prediction.

D. Random Forest

Random Forest may be a widely used supervised machine learning technique. It's a type of ensemble learning that's used to solve classification and regression problems. It generates multiple decision trees and analyzes them for making a decision. It is mainly used for classification problems. Random Forest algorithm can handle datasets containing continuous and categorical variables both. For regression problems, continuous variables are employed, while categorical variables are used for classification problems. Because it integrates several models, Random Forest is an ensemble algorithm. Here, a collection of

models is used rather than an individual model to make prediction easier. In the Random Forest algorithm, at first, some attributes are selected randomly from the data set. Then decision trees are constructed separately for each attribute. Each decision tree generates individual outputs using the decision tree method. And finally based on the maximum value of the decision tree, it generates the final output, Fig. 3.

As in comparison to other algorithms, it takes less time to train the dataset. When a massive quantity of the facts is missing, it can nevertheless preserve accuracy. It can predict output with proper accuracy, and it runs successfully despite a large dataset.

E. Support Vector Machine

Support Vector Machine is a supervised machine learning technique for categorization that analyzes data. It was developed by Vladimir Vapnik with colleagues in years 1992, 1993, 1995, 1997 [13]. Support Vector Machines (SVM) is a rapid and reliable classification method that works nicely with little amount records. Support Vector Machine is different from different classification algorithms because it chooses the decision boundary which maximizes the space from the nearest data factors. It is suitable for classification problems. The SVM algorithm's challenge is to discover a hyperplane in an N-dimensional space that exactly classifies the data factors. In the feature space, the SVM reveals the hyperplane which differentiates between the lessons [14]. For an SVM model, the data points which are trained, are segregated by a margin that belongs to a separate class. Then test data points are mapped into the same region to determine which side of the margin they will land on, Fig. 4.

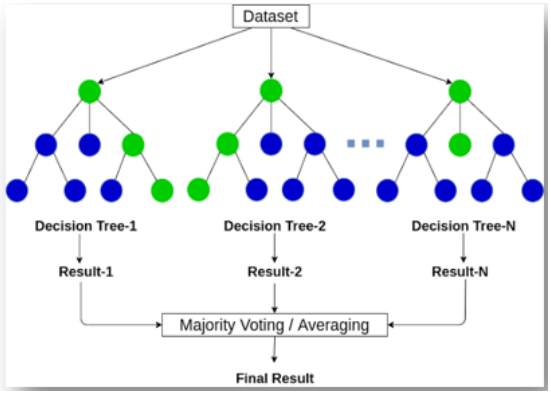


Fig. 3. Random Forest.

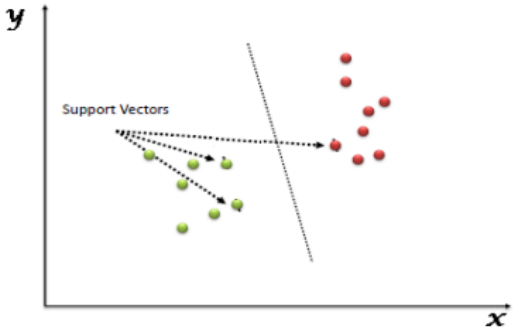


Fig. 4. Support Vector Machine.

In situations with a lot of dimensions, SVM works well. For the decision functions, several kernel functions can be given, as well as unique kernels. It saves memory by using a subset of training factors named support vectors in the selection feature.

V. Result and Evaluation

We employed five distinct types of classification algorithms to predict heart disease in this procedure. After preparing the data, we ran it through various categorization algorithms to see how well it performed. For heart disease prediction, we employed the K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine algorithms. We evaluated each algorithm's performance using accuracy, precision, recall, and f1 score values. We learned from our experiment that Naive Bayes performed the best of all algorithms, with an accuracy of 83.96 percent. Support Vector Machine performed admirably, with an accuracy of 84.08 percent, practically identical to Naive Bayes. Although it has higher accuracy than Naive Bayes, it performs poorly in terms of precision, recall, and F1 scores. The following "Table I" is a table of our experiment's performance measure:

TABLE I. Performance of Various Algorithms

Algorithms	Accuracy	Precision	Recall	F1- Score
KNN	81.13%	76%	81%	78%
Naive Bayes	83.96%	81%	84%	81%
Decision Tree	73.70%	74%	74%	74%
Random Forest	83.02%	76%	83%	78%
SVM	84.08%	71%	84%	77%

The performance of different algorithms is also represented through a bar chart, which is given in Fig. 5.

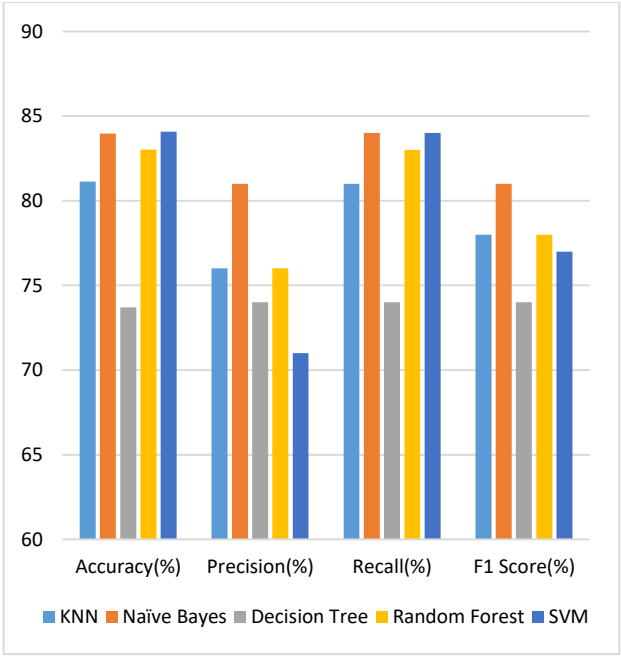


Fig. 5. Bar Chart of Performance for different Algorithms.

VI. Conclusion and Future Work

Machine learning can properly predict and guide treatment for heart disease, but models that include social determinants of health capture risk and outcomes for a wider range of people. A correct prediction of heart disease can save a person's life, while an incorrect prediction can be fatal. As can be seen from the preceding assertions, there is a lot of potential for applying various machine learning algorithms to predict cardiac disease. Our research aims to assess the performance of various machine learning algorithms and forecast which algorithm would perform better in this scenario. We've collected raw datasets, pre-processed them, and tested them for making a prediction. Some algorithms performed best or some performed worst in some cases. Naive Bayes performed the best accuracy for our dataset. Support Vector Machine algorithm also performed well but in comparison to Naive Bayes algorithm, its outcome was poor. Here, the Decision tree performed poorly in some cases. Random Forest also fared well since it used many Decision Trees to overcome the problem of overfitting. For our dataset, Naive Bayes performed well which can be used for predicting heart disease. By using these techniques for detecting heart disease, millions of lives can be saved.

The systems we employed in this work performed well in terms of predicting cardiac disease but still, there are some limitations in our research including limitation of processing power, time limit available for this research. Future research is needed to deal with high-dimensional data and overfitting. This document can serve as a starting point for learning how to anticipate cardiac disease, and it can be expanded to a more advanced level.

Acknowledgment

All thanks are due to Allah SWT, who created us and elevated us to the highest rank among his creations. First of all, I would like to admit my thanks to Allah SWT for enabling me to perform this thesis successfully. I'd like to thank my respected supervisor from the bottom of my heart, S. M. Hasan Sazzad Iqbal, Assistant Professor, Department of Computer Science and Engineering (CSE), Pabna University of Science Technology (PUST), for his scholastic supervision, valuable guidance, adequate encouragement and helpful discussion throughout the progress of this work. I owe him a debt of gratitude for enabling me to do this research under his supervision.

Finally, I owe a great deal to my family members, particularly my parents, as well as all of my friends and well-wishers for their encouragement and support.

References

[1] Rajpal, N. Decision Support System for Heart Di Decision Support System for Heart Disease Diagnosis Using Neural Network Niti Guru* Anil Dahiya.

[2] Sultana, M., Haider, A., & Uddin, M. S. (2016, September). Analysis of data mining techniques for heart disease prediction. In 2016 3rd international conference on electrical engineering and information communication technology (ICEEICT) (pp. 1-5). IEEE.

[3] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85-94.

[4] Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 36(4), 8616-8630.

[5] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems (IJAIS), 3(7), 25-30.

[6] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

[7] Alotaibi, F. S. (2019). Implementation of a machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 10(6), 261-268.

[8] Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.

[9] Rajathi, S., & Radhamani, G. (2016, March). Prediction and analysis of Rheumatic heart disease using kNN classification with ACO. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 68-73). IEEE.

[10] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 204-207). IEEE.

[11] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.

[12] Nahiduzzaman, M., Nayeem, M. J., Ahmed, M. T., & Zaman, M. S. U. (2019, December). Prediction of heart disease using multi-layer perceptron neural network and support vector machine. In 2019 4th International conference on electrical information and communication technology (EICT) (pp. 1-6). IEEE.

[13] Kannan, R., & Vasanthi, V. (2019). Machine learning algorithms with ROC curves for predicting and diagnosing heart disease. In Soft Computing and medical bioinformatics (pp. 63-72). Springer, Singapore.

[14] Schuldt, C., Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 3, pp. 32-36). IEEE.

Implementation of Modified Wiener Filtering in Frequency Domain in Speech Enhancement

Mr.C.Ramesh Kumar[1], Dr. M.P.Chitra[2]
Research Scholar Sathyabama Institute of Science and Technology[1]
Department of Electronics and Communication Engineering[1]
Assistant Professor, Panimalar Engineering College, Chennai[1]
Professor, Department of Electronics and Communication Engineering[2]
Panimalar Institute of Technology, Chennai, Tamil Nadu, India[2]

***Abstract*—The most common complaint about Digital Hearing Aids is feedback noise. Many attempts have been undertaken in recent years to successfully reduce feedback noise. A wiener filter, which calculates the wiener gain using before and after filtering SNR, is one technique to reduce background noise. Modified Noise Reduction Method (MNRM), a new way for reducing feedback noise Reduction, is presented in this work. In the Modified Noise Reduction Strategy, the advantages of a wiener filter are merged with a decision-directed approach and a twin-stage noise suppression technique The Modified Noise Reduction method can reduce the noise more successfully, according to comprehensive MATLAB programming, investigation, and findings analysis. After being modelled in MATLAB for seven distinct noise types, the SNR of the two architectures is compared.**

Keywords*—Digital hearing aid; least mean square value; noise reduction method; power spectral density

I. INTRODUCTION

The main purposes of DHA is to give frequency dependent amplification to persons who are deaf or hard of hearing. The DHA may also filter background noise, compress dynamic range, and remove feedback because it has an unit for external voice processing. Because digital hearing aids require frequency dependent amplification, sub band domain signal processing is a preferable option. When compared to the block transform, the lapped transform produces a bigger rejection of side lobes [22]. DFT, DCT, and MLT, AMLT, among others, are examples of block transformation techniques, whereas MLT, AMLT, and other lapped transformation techniques are examples of lapped transformation techniques [17].

DHA feedback can be cancelled using the LMS or NLMS algorithms. Because the NLMS algorithm has variable step size control, it has a better steady state behavior than the LMS algorithm. The two solutions proposed outperform Speech improvement using a decision-driven approach while retaining the decision-driven approach's advantages. The advantages of contemporary sub band domain digital hearing aid systems were explored by Ashutosh et al (2011) [1,2]. Easy gain adjustments in each sub band, very quick convergence of adaptive filters, and calculation savings are just a few of the benefits.

In today's sub band domain Digital Hearing Aids, the wiener filter is employed to reduce noise [3]. Feedback elimination is done using the NLMS approach. NLMS algorithm enhances active noise reduction, resulting in more consistent gain. To suppress feedback, adaptive filters in modern sub band DHA use block transformations like as DFT, GDFT. The adaptive filter techniques LMS and NLMS are used to eliminate the feedback noise by varying the filter coefficients [6]. Error e(n) provides the smallest term, the filter co-efficient is thought to be the best The co-efficient of these filters are utilised to predict constantly in the forward path and cancel the FB path [4,5]. Ear may be harmed if the increased speech volume surpasses the upper thresholds of hearing, hence DRC should be included in the hearing aid to avoid painful listening. Background noise suppression is the basic function of speech enhancement in DHA. Hearing aid users will benefit from improved unwanted sound suppression techniques that will improve performance and provide a more pleasant listening experience.

Multichannel DRC with decoded Digital Frequency Warping to eliminate noise, the DFW filter was replaced by an all-pass filter.[9] To compensate for hearing loss in digital hearing aids, the filter bank channel gains must be changeable over a wide dynamic range[10,11].In Biological basic DHA contain two channel ,the first channel includes a directional unit for receiving the acoustic input signal and providing a directional signal; a correlative unit for receiving the directional signal and providing a noise reduced signal [12,13].Compare to 16 point TAP wiener filter,32 point TAP wiener filter cancels 17db noise [14]. A delay-based NLMS technique is used to update the two adaptive filters [15,16]. Internal and external Acoustic feedback noise reduction using DSP Processor. Various technological and acoustic modification approaches are available to suppress or lesser auditory feedback. [18,19]. The effect of setting a hearing aid's gain control to a level just below that required to cause audible oscillation was explored [23,24]. Multi-channel compression techniques provide a realistic way to transfer speech signals' large dynamic range onto the reduced dynamic range of hearing-impaired listeners [25]. In DHA to regulate the gain value with different frequency and improve the SN [26]. Preprocessing filter algorithm used to High SNR with low noise speech signal produced in industry environment [27].

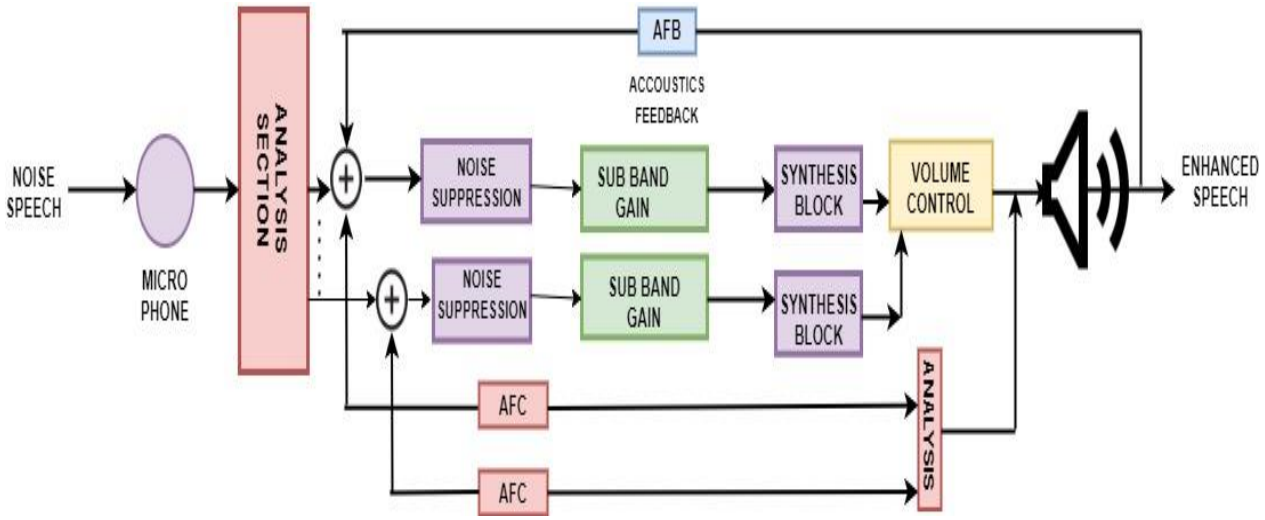


Fig. 1. Functional Diagram of Improved Sub Band Domain DHA.

A Wiener filter is used to reduce noise in most improved sub band domain digital hearing aids [7,8]. A block schematic of a modern sub band domain DHA is shown in Fig. 1. It works with the block transform's created subbands. As input to the DHA, the feedback passage signal from the AFB channel is combined with the noisy speech. Signal processing is used for hearing-impaired correction, noise elimination, AFB elimination, and DRC [20,21]. The analysis portion generates the appropriate subbands, which are then reproduced in the synthesis section.

This paper's structure is as follows: The first section contains a brief introduction. The Modified Subband Domain Digital Hearing Aid is shown in Section II. Sections A and B detail the Wiener Filter and MNRM algorithms. Section III concludes with the results, debates, and conclusions.

II. Traditional Wiener Approach

The Wiener filter is a well-known signal augmentation method that has been applied to a variety of applications. The Wiener filter's basic concept is to separate noise eliminated original speech from the noise mixed speech. Approximation is made by minimising the Mean. MSE between the desired signal S(n) and the error signal e(n). As a result of this optimization's frequency domain solution, the following filter transfer function is obtained.

$$H(W) = \frac{P_S(W)}{P_S(W) + P_V(W)} \quad (1)$$

Ps(W) – PSD of noiseless speech and Pv(W) -PSD of noise added Speech.

The SNR is defined as follows by [13]:

$$H(W) = \frac{P_S(W)}{P_V(W)} \quad (2)$$

H (W) can be Expressed in wiener Filter as follows:

$$\text{H}(W) = \left[1 + \frac{1}{Signal\ to\ Noise\ Ratio}\right]^{-1} \quad (3)$$

The Wiener filter has the disadvantage of having a constant frequency measures at every frequency and requiring calculation of the clean signal and noise power spectral density before filtering. This section shows how to create an AWF that uses the changing local statistics of the speech signal. The proposed technique is depicted in Fig. 1 as a block diagram. In this procedure, the expected mean value of speech signal μ_x and Var $\sigma_x{}^2$ are employed With a variance of σ_x, the added noise v(n) is said to have μ_x =0 and a white nature. As a result, we may take the Power Spectrum $P_V(w)$ as follows:

$$P_V(w) = \sigma_V{}^2 \quad (4)$$

A brief part of a voice transmission in which the input x(n) is regarded stationary.

$$\text{x(n)} = \mu_x + \sigma_x\, w(n) \quad (5)$$

The Mean μ_x and SD of σ_x are represented by, A noise with a unit variance is called a unit variance noise w (n)The Wiener filter transfers inside this tiny section of speech. The following functions can be used to approximate the function:

$$H(w) = \frac{P_S(w)}{P_S(w) + P_V(w)} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_V^2} \quad (6)$$

Eq. (12) may be used to compute the Wiener filter's impulse response because H (ω) is constant for this little portion of speech:

$$H(w) = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_V^2}\,\delta(n) \quad (7)$$

Eq. (13) can be used to represent the improved voice signal s(n) in this local field.

$$\hat{S}\text{(n)} = \mu_x + (\text{x(n)} - \mu_x) * \frac{\sigma_S^2}{\sigma_S^2 + \sigma_V^2}\,\delta(n) \quad (8)$$

$$= \mu_x + \frac{\sigma_S^2}{\sigma_S^2 + \sigma_V^2}(\text{x(n)} - \mu_x) \quad (9)$$

We can say: if we update μ_x

$$\hat{S}\text{(n)} = \mu_x + \frac{\sigma_S^2}{\sigma_S^2 + \sigma_V^2}(\text{x(n)} - \mu_x(n)) \quad (10)$$

Eq. (10) modifies the mean μ_x (n) and (x(n) $-$ μ_x (n)) independently from section to section before combining the results. The result signal s(n) will be predominantly owing to x(n) if σ_S^2 is significantly larger than σ_V^2, and x(n) cannot be changed. When σ_S^2 is less than σ_V^2, the filtering effect is applied. When μ_V is zero, you'll notice that μ_x is the same as μ_S. As a result, we may calculate μ_x(n) in Eq. (11) using x(n) by.

$$\widehat{\mu_S}(n) = \widehat{\mu_x}(n) = \frac{1}{2M+1} \sum_{K=n-M}^{n+M} x\,(k) \tag{11}$$

$$\sigma_S^2\,(n) = \begin{cases} \sigma_x^2\,(n) - \sigma_V^2\,(n); if\ \sigma_x^2\,(n) > \sigma_V^2\,(n) \\ 0; Other\ wise \end{cases} \tag{12}$$

(2M + 1) samples were used in the estimation in the brief part. We need to calculate the speech by.

$$\sigma_S^2 = \sigma_S^2 - \sigma_V^2 \tag{13}$$

calculate the signal variance σ_S^2 as follows from x(n):

$$\sigma_x^2\,(n) = \frac{1}{2M+1} \sum_{K=n-M}^{n+M} (x\,(k) - \widehat{\mu_x}\,(n))^2 \tag{14}$$

A. Conventional Noise Suppression Technique using Wiener Filter

This filter is used to reduce background noise. Wiener filter reduces noise by increasing the gain by the noisy voice input. The operations involved in wiener filtering noise reduction are depicted in Fig. 2. A noisy speech signal is fed into the Wiener filter, which separates it into N frames. FFT is used to convert time domain to frequency domain output. To achieve the high throughput and reduce the area instead of conventional FFT R2^{2}SDF FFT is proposed in this work. PSD is calculated using noise characteristics. The initial silent area is used to calculate the noise characteristics of voice communications. The wiener gain is increased by the noisy voice signal in the frequency nature after it has been calculated.

The overall gain of the noisy voice signal is reduced to achieve noise reduction. Because the short temporal energy of a noisy signal is smaller than that of a speech signal, the noise signal is suppressed even after the noise suppressed speech signal is enhanced. Linear Filtering can be achieved by using Overlap Add Method.

Overlap Add Method (Linear Filtering)

Steps to compute overlap add:

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Divide the long sequence signal into L-length small sequence.
- Insert zeros into this small sequence, increasing the size of the new sequence to m + n, since m +n = 2^n.
- Calculate the FFT of the full sequence, including the padding zeros.
- To resynthesize the signal, utilise IFFT and overlapping and adding the result.

B. Proposed Modified Noise Reduction Method (MNRM)

The modified noise reduction method operates in the frequency field and combines the qualities of a Wiener filter based on a decision-directed approach with a twin-stage noise suppression technique Fig. 3 depicts the activities involved in the Modified Noise Reduction Method (MNRM) for digital hearing aids. A noisy speech signal is fed into the INST filter, which separates it into N frames. FFT is used to convert time domain to frequency domain output. To achieve the high throughput and reduce the area instead of conventional FFT R2^{2}SDF FFT is proposed in this work. Noise characteristics are calculated from the first few frames. PSD is calculated using noise characteristics. The initial gain is determined in the same way as the Wiener filter gain after the pre and Post SNRs have been calculated. The cost of the MNRM is minimised by treating the posteriori and priory signal to noise ratio as the wiener filter gain, because the DHA must be operated in real time. N frames are created from the signal.

Because the MNRM's computing complexity is comparable to that of the wiener filter for noise removal, the proposed MNRM method can operate in real time, providing a better and more comfortable listening experience for hearing impaired people.

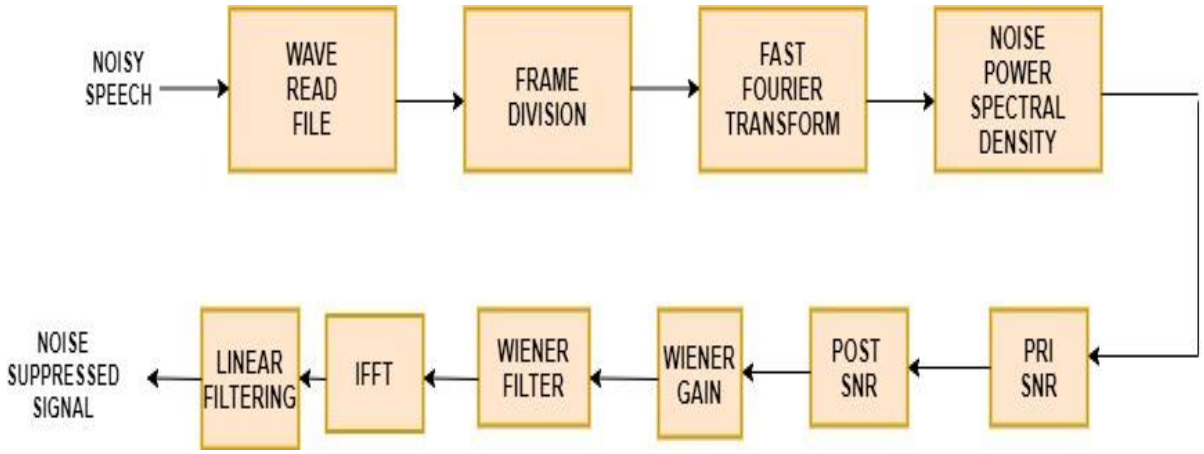


Fig. 2. Noise Reduction by Wiener Filter in DHA.

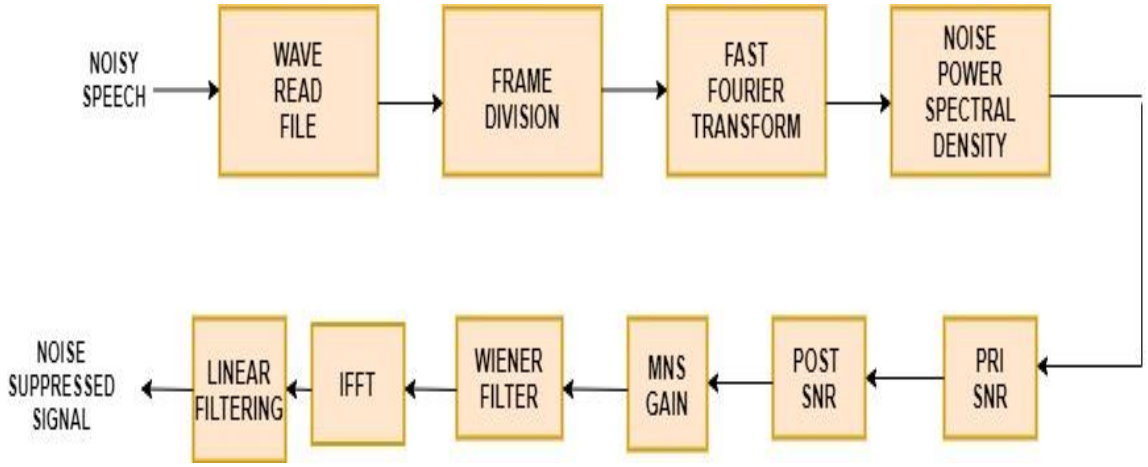


Fig. 3. Modified Noise Reduction Method in DHA.

III. Experimental Results and Discussion

In MATLAB, we replicated a modern DHA with sub band domain with Compensation for hearing loss, DRC, noise reduction, and feedback cancellation for comparison. Four sub bands (S1) and eight sub bands (S2) are used in the two sets of SP in the sub band field. The research was undertaken performed on a variety of background noisy voice signals from the NOIZEUS database. The signals used to calculate the test signals the DHA performance are 8 sets of noisy speech sounds at varying SNRs such as 0dB, 5dB, 10dB, and 15dB. Fig. 6 depicts the performance of MNRM and wiener filtering approaches for noise suppression in DHA applications. We tried two ways using the identical type of loud voice signals for both methods for the sake of performance testing.

Fig. 4 shows the noise elimination of wiener filtering. Left side represents Time domain signal for clean speech, Noisy speech with street Noise and Enhanced Speech using Wiener and right side represent the Spectrogram for Clear, Noisy and Enhanced speech using Wiener. Fig. 5 shows the noise elimination of MNRM filter. Left side represents Time domain signal for clean speech, Noisy speech with street Noise and Enhanced Speech using MNRM and right side represent the Spectrogram for Clear, Noisy and Enhanced speech using MNRM filter. The evaluation of the result is computed by adding the absolute amplitude of the improved voice signal at the outcomes of the DHA. The figure's analysis in some trails, Fig. 6 reveals that the MNRM has a slight advantage over the wiener filter in terms of noise suppression.

The SNR obtained with various Ambient noises is compared for both approaches are shown in Table I.

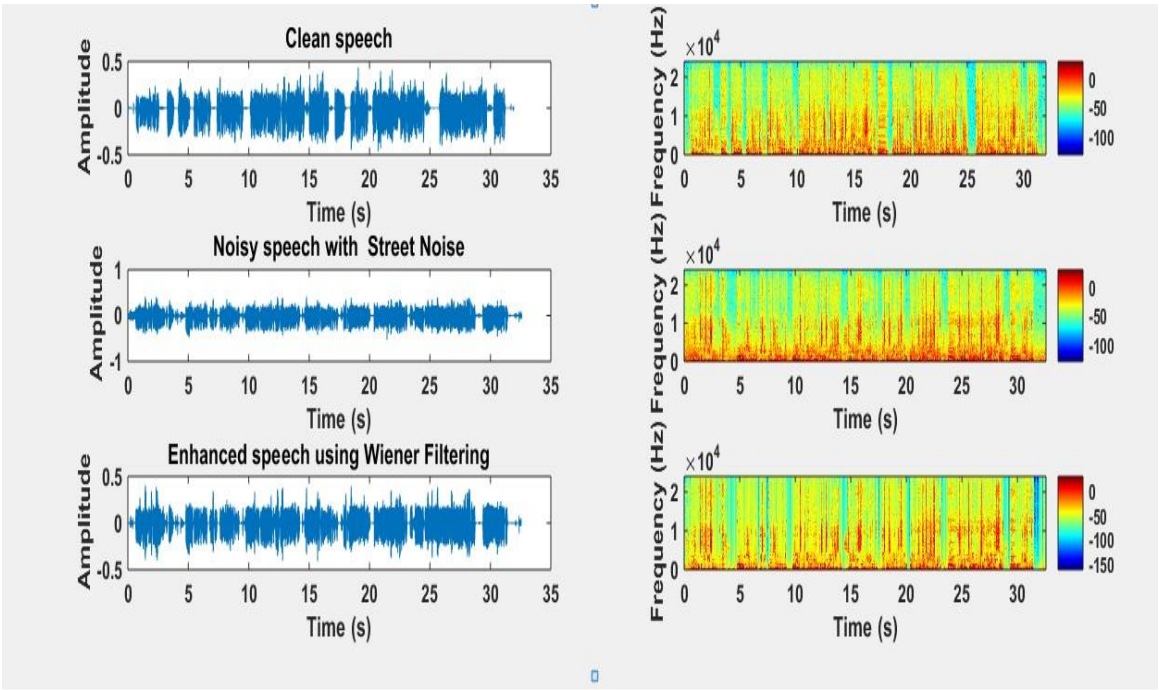


Fig. 4. Left Side represents Time Domain Signal for Clean Speech, Noisy Speech with Street Noise and Enhanced Speech using Wiener and Right Side represent the Spectrogram for Clear, Noisy and Enhanced Speech using Wiener.

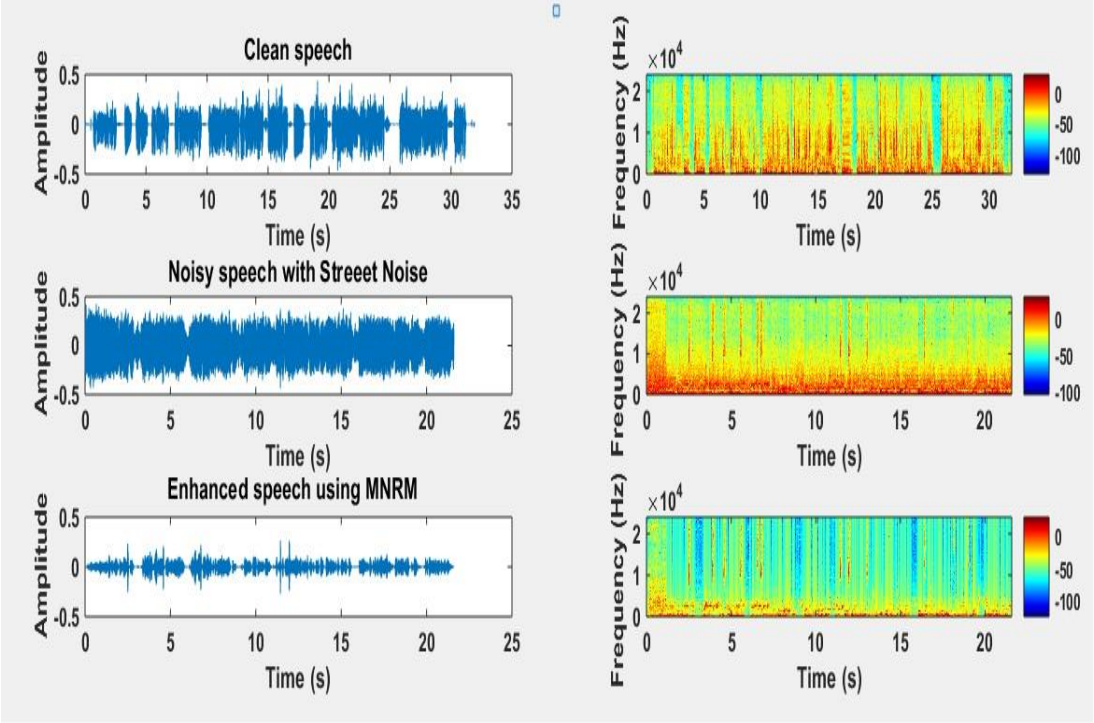


Fig. 5. Left Side represents Time Domain Signal for Clean Speech, Noisy Speech with Street Noise and Enhanced Speech using MNRM and Right Side represent the Spectrogram for Clear, Noisy and Enhanced Speech using MNRM.

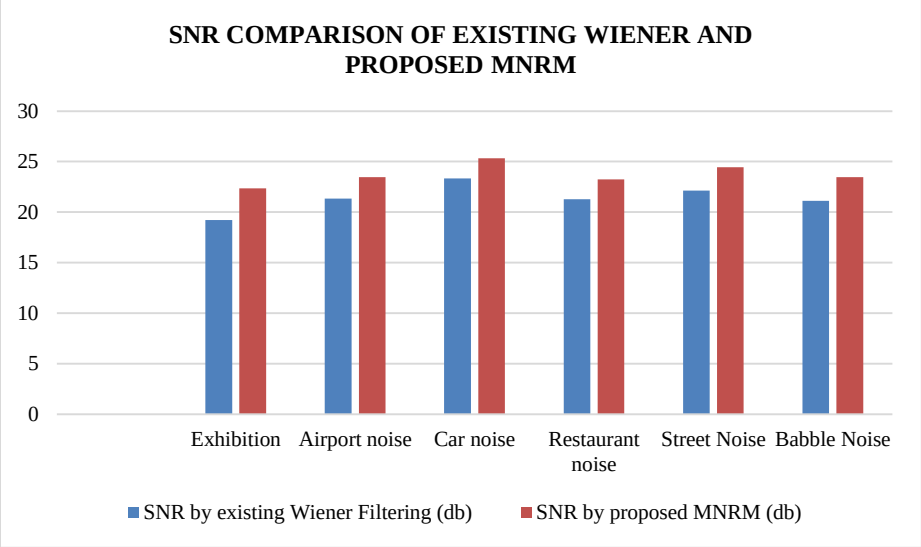


Fig. 6. SNR Comparison of Wiener and MNST Filters.

TABLE I. COMPARING SNR OF WIENER FILTERING AND MNRM

Various Ambient noises	SNR by existing Wiener Filtering (db)	SNR by proposed MNRM (db)
Exhibition	19.234	22.345
Airport noise	21.345	23.456
Car noise	23.345	25.345
Restaurant noise	21.267	23.234
Street Noise	22.145	24.456
Babble Noise	21.134	23.456

IV. CONCLUSION

This research developed a novel approach for suppressing background noise in DHA via signal processing. In Matlab simulations, the performance of MNRM is compared to that of the wiener filter. In this work concluded with seven different Ambient noises (Exhibition, Airport, Car, Restaurant, Street, Babble) compare with Wiener and MNRM filters. Based on the comparison these two filters for the SNR ratio for those different noises increased with our proposed model of MNST filter. Approximately 14 to 15 % of SNR ratio increased by using with proposed model. The MNST can be employed in modern subband domain DHA for noise reduction, as shown in Matlab. The user of hearing aids will benefit from this, with a better and more comfortable listening experience.

REFERENCES

[1] Ashutosh Pandey, V. John Mathews,"Low-Delay Signal Processing for Digital Hearing Aids", IEEE Trans. Audio,Speech and Language Processing, 19, pp. 699-710, 2011.

[2] Hellgren J, "Analysis of feedback cancellation in hearing aids with filtered-X LMS and the direct method of closed loop identification", IEEE Trans. Speech Audio Process., 10, pp.119–131, 2002.

[3] Philips, C. Loizou, "Speech Enhancement Theory and Practice", CRC Press.

[4] Siqueira M. G., Alwan A.,"Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids", IEEE, Trans. Speech Audio Process., 8, pp.443–453, 2000.

[5] Stone M. A., Moore B. C. J.,"Tolerable hearing aids delays.II: Estimation of limits imposed during speech production",J. Amer. Academics of Audiology, 11, pp. 325–338, 2002.

[6] Sunitha S. L., Udayashankara V.,"Fast Factored DCTLMS Speech Enhancement for Performance Enhancement of Digital Hearing Aid", World Academy of Science, Engineering and Technology, 10, pp. 253-257, 2005.

[7] Wyrsch S., Kaelin A.,"Subband signal processing for hearingaids", In Proc. IEEE Int. Symp. Circuits Syst., 3, pp. 29–32, 1999.

[8] S.Wyrsch and A. Kaelin, "Subband signal processing for hearing aids,"in *Proc. IEEE Int. Symp. Circuits Syst.*, Orlando, FL, Jul. 1999, vol. 3,pp. 29–32.

[9] K. M. Kates and K. H. Arehart, "Multichannel dynamic-range compression using digital frequency warping," *EURASIP J. Appl. Signal Process.*, vol. 18, no. 1, pp. 3003–3014, Jan. 2005.

[10] R. Brennan and T. Schneider, "A flexible filter bank structure for extensive signal manipulations in digital hearing aids," in *Proc. IEEE Int. Symp. Circuits Syst.*, Monterey, CA, Jun. 1998, vol. 6, pp. 569–572.

[11] M. Harteneck, S. Weiss, and R. W. Stewart, "Design of near perfect reconstruction oversampled filter banks for subband adaptive filters", *IEEE Trans. Circuits Syst.*, vol. 46, no. 8, pp. 1081–1086, Aug. 1999.

[12] M. B. Sachs, I. C. Bruce, R. L. Miller, and E. D. Young, "Biological basis of hearing-aid design," *Ann. Biomed. Eng.*, vol. 30, no. 2, pp. 157–168, Feb. 2002.

[13] D. Bustamante, T.Worrall, and M. Williamson, "Measurement of adaptive suppression of acoustic feedback in hearing aids," in *Proc. 1989IEEE ICASSP*, 1989, pp. 2017–2020.

[14] J. Kates, "Feedback cancellation in hearing aids: Results from a computer simulation," *IEEE Trans. Signal Processing*, vol. 39, pp. 553–562,1991.

[15] J. Maxwell and P. Zurek, "Reducing acoustic feedback in hearing aids,"*IEEE Trans. Speech Audio Processing*, vol. 4, pp. 304–313, July 1995.

[16] P. Estermann and A. Kaelin, "Feedback cancellation in hearing aids Results from using frequency-domain adaptive filters," in *Proc. 1994 IEEE ISCAS*, 1994, pp. 257–260.

[17] A. Oppenheim and R. Schaffer, *Discrete-Time Signal Processing*.Englewood-Cliffs, NJ: Prentice-Hall, 1989.

[18] Agnew, J., 1993. Application of a notch filter to reduce acoustic feedback. Hearing J. 46, 37–43.

[19] Bustamante, D.K., Worral, T.L., Williamson, M.J., 1989.Measurement of adaptive suppression of acoustic feedbackin hearing aids. In: Proc. IEEE ICASSP-89, pp. 2017–2020.

[20] Chi, H.F., 1999. Adaptive feedback cancellation for hearing aids: Theories, algorithms, computations, and systems. Ph.D. dissertation, Department of Electrical Engineering, University of California, Los Angeles.

[21] Chi, H.F., Gao, S.X., Soli, S.D., 1999. A novel approach ofadaptive feedback cancellation for hearing aids. In: Proc.IEEE ISCAS-99, pp. 187–190.

[22] Cox, R.M., 1982. Combined effects of earmold vents and suboscillatory feedback on hearing aid frequency response.Ear Hearing 3, 12–17.

[23] Dillon, H., 1991. Allowing for real ear venting effects when selecting the coupler gain of hearing aids. Ear Hearing 12,406–416.

[24] B. Gold, "Robust speech processing," M.I.T. Lincoln Lab., Tech. Note 1976-6, DDC AD-A012 P99/0, Jan. 27, 1976.

[25] Schneider, **T.,** Brennan R.L., "A Multichannel Compression Strategy for a Digital Hearing Aid," *Proc. ICASSP-97*, Munich, Germany, pp. 411-415.

[26] B. Saha, s. Khan, c. Shahnaz, s. A. Fattah, m. T. Islam and a. I. Khan, "configurable digital hearing aid system with reduction of noise for speech enhancement using spectral subtraction method and frequency dependent amplification," tencon 2018 - 2018 ieee region 10 conference 2018, pp. 0735-0740, doi: 10.1109/tencon.2018.8650450.

[27] N. M. Yunus, n. A. Noor affande, r. M. Ramli, a. O. A. Noor and s. A. Samad, "preprocessing noise reduction for assistive listening system," 2021 1st international conference on electronic and electrical engineering and intelligent system (ice3is), 2021, pp. 114-119, doi: 10.1109/ice3is54102.2021.9649710.

A Framework for Integrating the Distributed Hash Table (DHT) with an Enhanced Bloom's Filter in MANET

Ms. Renisha P Salim
Research Scholar
Bharathiyar University and Assistant Professor
Providence College of Engineering, India

Dr. Rajesh R
Associate Professor
Department of Computer Science
Christ (Deemed to be University), India

***Abstract*—MANET, a self-organizing, infrastructure-less, wireless network is a fast-growing technology in day-to-day life. There is a rapid growth in the area of mobile computing due to the extent of economical and huge availability of wireless devices which leads to the extensive analysis of the mobile ad-hoc network. It consists of the collection of wireless dynamic nodes. Due to this dynamic nature, the routing of packets in the MANET is a complex one. The integration of distributed hash table (DHT) in MANET is performed to enhance the overlay of routing. The node status updating in the centralized hash table creates the storage overhead. The bloom filter is a data structure that is a space-effective randomized one but it allows the false-positive rates. However, this can be able to compensate for the issue of storage overhead in DHT (Distributed hash table). Hence, to overcome the storage overhead occurring in DHT, and reduce the false positives, the Bloom's filter is integrated with the DHT initially. Furthermore, the link stability is measured by the distance among mobile nodes. The optimal node selection should be done for the transmission of packets which is the lacking factor. If it fails to select the optimal path then the removal of malicious nodes may lead to the unwanted entry of nodes into the other clustering groups. Therefore, to solve this problem, the bloom's filter is modified for enhancing the link stability. The novelty of this proposed work is the integration of Bloom's filter with the Distributed Hash Table which provides good security on transmission data by removing false-positive errors and storage overhead.**

Keywords*—Mobile ad hoc network (MANET); distributed hash table (DHT); bloom's filter; link stability

I. INTRODUCTION

At present, Wireless communication shows a dynamic role in day-to-day life which leads to substantial growth in the mobile ad-hoc network (MANET). The MANETs are most commonly used in the area of military services, emergency services, sensing and gaming, education, the personal area network, and so on. It yields the benefits of centralized management of security, scalability, and good connectivity. It is the collection of various wireless dynamic nodes. Routing is the most significant method in the mobile ad-hoc network. The nodes that were present in the mobile ad-hoc network will perform the function of the router for transmitting the packets. Generally, the communication of the packet is done without any priory fixed network structure. Due to the dynamic nature of the mobile ad-hoc network, the routing of packets will be a complex one.

MANET is a kind of network with rapidly changing topology and are having a huge span with the capacity to connect hundreds [1] and thousands of nodes. The important feature of the mobile ad-hoc network is to identify the optimal path for the transmission of packets between the nodes. The nodes present in the mobile ad-hoc network are moving freely which affects its fixed infrastructure, hence it is necessary to introduce a dynamic network behavior.

The minimization of the network traffic is a significant part of constructing a MANET. For this, [2] a better QoS is essential to face this demand. Various mismatches [3] were identified in the construction of distributed hash tables like traffic overhead, high path stretching ratio, and long route. Also, in the traditional distributed hash table (DHT), each node of DHT will be proficient of maintain its table for the routing process. It also contains a list of communication links that are to be taken place without any awareness regarding the nodes of neighbor. The overlay of routing should be improved by integrating the distributed hash table with MANET.

The bloom filter is the randomized space-efficient structure of data for representing the [4] queries of support members. Conversely, the bloom filter is capable of allowing false positives. They are becoming very popular in the application of networks because of their concise size. The space-saving provided by this bloom filter will compensate for the issue of error occurrence probability.

Therefore, it is necessary to shrink the storage overhead and to diminish the false-positive rates of both the DHT and Bloom filters.

A. Problem Identification

The following problems are identified in the existing methodologies:

- The uncertainty of the route prediction is created due to the dynamic nature of the mobile nodes.
- The fault tolerance service is one of the major disputes in the MANET environment, provision of the optimal solution for this problem is necessary.

- Generally, link stability is restrained by the distance between the mobile nodes. But it fails to choose the optimal node for the transmission of data without the consideration of speed, mobility, and the direction of mobility.
- Removal of malicious nodes will cause the nodes to enter into the other clustering groups.

In the proposed system, the hash table is combined with the bloom filter and the modification of the bloom filter is done for resolving the issues of storage overhead, false-positive errors and to improve the link stability.

B. Objectives

The main intention of this proposed system is as follows:

- To reduce the false-positive rates by combining the blooms filter with DHT.
- To enhance the link stability equation for computing the optimal routing path through modifying bloom's filter approach.
- The attacker and position possibility is included for improving the link stability.
- To maintain the centralized hash table with the bloom's filter.

The remaining portion of this paper is schematized as follows: Section II demonstrates the conventional works associated with the DHT, Bloom's filter for MANET in the wireless sensor networks. Section III illuminates the proposed (DHT-MBF) a novel distributed hash table integrated with the modified bloom filter. Section IV exhibits the performance study of the proposed mechanism and in conclusion, Section V concludes the proposed work.

II. Related Work

This section deliberates the literature review of the routing protocols in MANET.

A Review was made on the routing protocols of the MANET in which the range of available routing protocols are discussed with their functionalities [5] which varies from protocols of early-stage to advanced protocols. It mainly focuses on developing and enhancing the MANET routing by Perkins. In general, ad-hoc networks offer a good potential in the circumstance at which access to the internet was not a chief requirement. Thus the evolution was made on the AODV protocol on analyzing the work depending upon the Multicast Ad-hoc on-demand Distance vector. However, this work needs some analysis on mobility-aware routing, Hierarchical routing, reliability-focused routing, and power-aware routing.

A survey was made on the distributed Hash table-dependent routing and the management [6] of data at the wireless sensor networks. The combination of the hash table over the wireless sensor network was made to manage the independent location of data and the identification of nodes. Various existing Hash table-based routing and the management of data protocols were described with their categories. Moreover, the detection of asymmetric link, bootstrapping and sensor dynamism was deliberated rarely which needs further analysis.

The analysis was made on the problems that were challenging in the mismatch among [7] resilience of overlay structure protocols and the physical networks. To overwhelm the difficulties of delay in average file discovery relay, the overhead of routing, high rate of average path-stretch, increased false-negative ratio, a distributed algorithm for exploiting the overlay and computing the subsequent logical identifier of the peer as defined in this work. However, there were some limitations like user anonymity, P2P partitioning of the network, load balancing, and free-riding issues.

The exploitation of 3-D structure was [8] made in MANET for the scalable routing is performed for avoiding the mismatch among the structure of logical identifier and the physical networks. This in turn preserves the issue of traffic overheads, the ratio of high path-stretch, and long routes. The approach of 3-D LIS was presented for managing the multi-paths at a destination node. The limitation of this work was Network merging and network partitioning.

A different methodology for the detection of service for MANETs was presented in which the protocols [9] of the service discovery efforts to overwhelm the incapability to abode the assets located at the network in which the node identity, its preceding knowledge, and also its ability is not accessible. The cross-layer approaches for the identification of service in MANET have been obtained to improve the procedure of discovery by the direct incorporation of the routing protocols. The novel service-oriented protocol for routing in MANETs was defined in this work. This method has improved success rates in relation to discovery and the throughput application in the densities of higher nodes.

An approach for maximizing the hash table throughput [10] at the network devices by combining the SRAM/DRAM Memory was performed. The Hash table is capable of forming a core component of a huge number of algorithms and network devices. It always requires a joint memory model as of its size at which some elements were kept in the slow memory and others were capable of storing in fast memory. The impacts of the memory speed difference with the choice of parameters were evaluated and the performances were traced. The employment of multiple-choice hashing was performed with the aid of combined memory. However, there were some limitations of this approach.

A QoS routing protocol depending on the link-state was presented which was established on the stability [11] of association for the mobile ad-hoc network. This approach was established to conserve the sustainable and stable path among the entire sets of nodes at a mobile ad-hoc network. The stability function was utilized as the main path for the selection criteria which was established on the degree of mobility in a node relation and its neighbor. For electing the constant and defensible MPR nodes and topologies this approach would be applied. The recompilation of MPR and routing tables were reduced by this mechanism. However, the metrics like packet loss and response time would be assured by applying this technique. Moreover, this approach needs some improvement

by integrating with the other protocols of routing like DSR and AODV by executing in the real-time application.

An effectual and accessible dual region-based management of mobility for the mobile [12] ad-hoc network was presented for attaining the management of the location of nodes at the ad-hoc network. However, there were not at all selfish nodes or malicious nodes at the MANET for disrupting the management of mobility. Hence, it was necessary to investigate the management of trust protocols to choose the trustworthy nodes by enhancing the activity of dual-region management of location in the MANETs.

The routing technique was the most stimulating one in the mobile ad-hoc network [13] for its dynamic topology. There exist several types of routing protocols that are quite complex to identify the suitable routing protocols to the circumstances of the network. This paper provides the depiction of various routing protocols and the difference among them. The features of the MANET network were categorized as follows: the communication was through the wireless type, the functions of hosts and routers were done by the nodes, and also it was a bandwidth-constrained one, and so on. However, there were some limitations in different routing protocols which was complex to select various routing protocols along with the situations. Hence, it was necessary to face these contests for future widespread use.

An efficient addressing [14] protocol for the auto-configuration of nodes at the ad-hoc networks was presented. The lightweight protocol for the configuration of the mobile ad-hoc network depends on the databases that were stored on the distributed address. From this approach, it was revealed that the address collisions were solved thereby reducing the control traffic on comparing traditional methodologies. The bloom filter was utilized in this approach. This approach was simple and diminishes the delay. However, there were some limitations in this kind of network.

A scheme of [15] congestion control in the case of a heterogeneous wireless ad-hoc network by using the self-adjust hybrid model was presented. The prediction of this heterogeneous wireless ad-hoc network was a complex one because of the occurrence of resource nodes in the distributed surroundings. The bloom filter was utilized for minimizing the error prediction which in turn reduces the fluctuations of load. The execution time was also increased. Anyhow this approach needs some improvement.

By using the [16] hierarchical bloom filter, the scalable content-based routing was presented for enabling the sharing, storing, and searching of information in the VANET. It offers a low rate of latency, ability to tackle mobility. This in turn offers a high completion rate as a result of the utilization of bloom's filter.

The addressing of [17] node replication was described and a different distributed hierarchical mechanism was introduced for the identification of node replication with the aid of the bloom filter approach. This provides an energy-efficient system in wireless networks. Moreover, this mechanism fails to validate the blooms filter through witnesses and in turn attain similar outcomes through identification.

The classification approach of the multi-label by using a large number of labels was [18] presented for the reduction of dimensional binary vectors. This scheme was robust to the inference complexity, sublinear training. The attitude was centered on the bloom filters mechanism. The computational difficulty was the shortcoming of this mechanism.

The optimization of traffic [19] in P2P centered on the congestion distance and DHT was stated which has been a major influence on the network of bearer type. The utilization of bandwidth networks has become a major problem for the user. The information collectors were assembled to the distributed hash table by the realization of information collectors and flow controllers. Therefore, by this scheme, the inter-autonomous system was reduced and the network traffic was diminished. To enhance the network traffic, this mechanism could be used by the other distributed system.

Abbas et al proposed a routing scheme [20] for limited configured devices. They uses on demand Ad-hoc distance vector algorithm for deriving the shortest path to the destination. Their model is validated in a simulated environment in Karachi and the experimental results claims that the performance is enhanced with minimal packet drops and constant throughput during the communication process. End-end communication delay is comparatively more which is a major drawback in their model.

Srilakshmi et al proposed a secure hybrid model for multipath routing [21] in the MANET. This work uses genetic algorithm with hill climbing which provides the optimal path. It uses prediction mechanism to choose the trust worth cluster heads and chooses alternate paths to reach the destination. It uses optimal energy for processing and is free from selective packet dropping attack.

III. Proposed Work

This section deliberates the proposed method used for the enhancement of link stability for computing the optimal routing path on modifying Bloom's filter. The integration of Bloom's filter with the Distributed Hash Table is performed to reduce the storage overhead and occurrence of false positives.

Fig. 1 describes the overall flow of the prosed system in which the ID-based public key is generated and neighbor nodes are discovered. After sharing the public key, the hash table is generated. Then the bloom filter sends and receives the message after the discovery of nodes and updating of the distributed hash table. The received hash value must be verified in the DHT. If it does not match the condition then it is considered as the malicious node in the blocked list or else the data communication is performed for analyzing the link.

A. Network Formation Module

In the network module formation, the generation of ID-based public keys is carried out initially. The mobile nodes in the network will dynamically set a path temporarily to transmit the packet within them.

The discovery of neighborhood nodes is performed after the public key generation which in turn shares the public key to the neighborhood node. In MANET, the nodes are responsible for selecting the subset of the neighbor nodes to transmit the

packet with minimum overhead. The development of a key management approach should be provided for the need to manage the routing table.

B. Generation of Hash Table

According to the ID-based public key that was created, the hash table H is generated. After that, the discovery of the route is carried out. The nodes that are located within one another's range will be communicated directly. The nodes that were located outside the ranges will be communicated through the other nodes for relaying the messages which should be performed with the routing protocols. Then the hash table is being updated for sending the message among nodes. The distributed hash table is updated according to the discovery of the route. The message that is to be transmitted should be coded or updated in the distributed hash table for security.

The routing of packets in MANET is a complex one because of its dynamic nature. Thus to improve the overlay of routing, the distributed hash table is integrated with the MANET. Each node in the traditional DHT maintains its routing table along with a list of links that are to be established for communication. Since they are unidirectional links it may unaware of neighbor nodes. From the neighbor list, the node chooses the routing path. The node status will be updated in the centralized portion of the hash table. This may lead to the storage overhead in the distributed hash table. It is necessary to reduce the storage overhead by integrating the bloom filter.

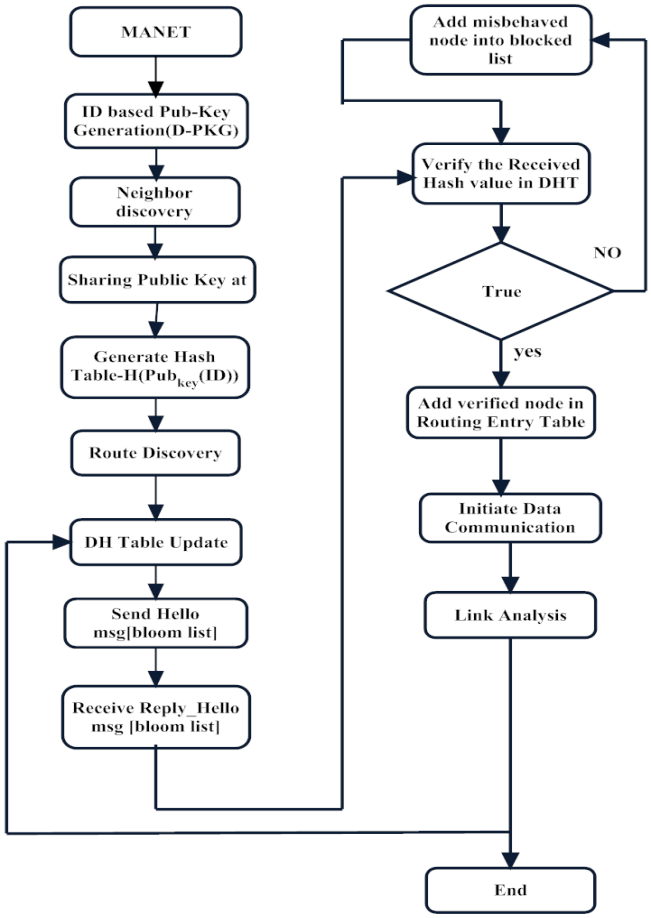


Fig. 1. Flow of the Proposed System.

The generation of the hash table was made by using node ID. The location and node ID are broadcasted and then the neighbor length is identified. The route is discovered by considering the source and destination id. The node should be sent by checking the path that is to be transmitted. This could be found using the own id of each node. For the next time of transmission, the packet should be transmitted through another hop. So the identification of hop is made to check if it is not the same hop through which the packet is already sent. Finally, the hash table is updated with the node status, the number of hops available for transmission, and the information regarding each hop.

Algorithm: Generation of Hash table using node ID

Step 1: Broadcasting the node id and location
Step 2: $for\ i:\ 1: length\ of\ neighbor$
$p \leftarrow Hash_{key}(Id) \otimes m$
$DHTable.id[i, Pub_{key}] \leftarrow 1$
$DHTable.id[i, H_i(p)] \leftarrow 1$
Step 3: $if\ source\ =\ Own(id)\ then$
$ttl \leftarrow MAXTTL + 1$
$Next\ hop\ of\ n(id)$
$list\ node \rightarrow \{id\} iii$
$Send\_msg\ (\{\ RwREQ;\ source\ ;\ dest\ ;\ list\ node;\ ttl\ \},$
$UNICAST, next\ hop)$
$Set\_timeout\ (\{\ dest\ \}, WAITTIME)$
$end\ if$
Step 4:
$If\ Receive\_msg(RwReQ, Source, dest, list\_node, ttl)\ then$
$local\_heat\_level \leftarrow Search(GTAB, dest)$
$if\ local\_heat\_level ==\ NONE\ and\ ttl > 0\ then$

$next\_hop \leftarrow choose\_random\_neighbour()$
$ttl \leftarrow ttl - 1$
$list\_node \leftarrow list\_node\ U\ myID$
$Send_{msg}(\{RwREQ, source_{dest}, list_{node}), ttl\}, UNICAST,$
$nexthop)$
$else\ if\ local\_heat\_level \neq\ NONE\ then$
$Send\_msg\ (\{FoHEAT, source, dest, list\_node,$
$local\_heat\_level\}, BROADCAST)$
$endif$
$endif$
Step 5:
$if\ Receive_{msg(RoRep, soure, dest, prev_{node}, list_{node})}$
$then$
$next\_hop \leftarrow list\_node$
$if\ source == myID\ then$
$ADD\_routing\_entry\ (prev\_node, destination)$
$Forward\_data\_packet\ (next\_hop)$
$else$
$next\_hop \leftarrow Last(list_{node})$
$list\_node \leftarrow Remove\_last(list\_node)$
$Add\_routing\_entry(prev_{node}, destination)$
$Send\_msg(\{RoREP, source, dest, prev_{node}, list_{node}\},$
$UNICAST, next\_hop$
$endif$
$endif$

C. Bloom's Filter

The data or message will be transmitted after the updating of the message in the distributed hash table. The message is transmitted with the help of the bloom list and is received with the same bloom filter. A bloom's filter is a randomized space-efficient structure of data for representing the set and supporting membership queries. However, it may allow some false positives they may capable of offering some space-efficient environment which compensates for the storage overhead occurring in DHT (Distributed hash table).

Bloom filters offer a tool for the identification of member set function which may include N number of array bits and various hash functions of K. On installing the element, then the system is capable of sending an element to the various hash functions K and also sets the bits in an array by using the output values of the hash function. The system will check array bits that are corresponding to the hash function output values once the identification of the membership status of the element is set. After, the setting of all array bits to the respective hash function output values, then the system estimates the element as a member of all sets.

D. Data Transmission

After, the transmission of the message by bloom filter the verification is done in the received hash table to that of the DHT to check whether the sent and received messages are correctly matched. If this condition is satisfied then the verified node is added in the entry table of routing. If it fails to verify the condition, then it will be added in the blocked list as a misbehaved node which is then re-verified.

Finally, the data communication is initiated and the link stability is analyzed. Generally, link stability is measured by estimating the distance among the mobile nodes. The establishment of the connection is affected by the occurrence of a removed malicious node in another cluster. Link stability needs an optimal path for routing during the transmission of data. Depending on the real-time application, optimal path identification is an important process. To enhance the link stability, bloom's filter is modified. Hence, for the selection of optimal nodes, the reduction of false-positive rates and storage overhead is considered as a significant part.

Therefore, the integration of the bloom filter with the distributed hash table is much important to compute the optimal path, the reduction of storage overhead, and false-positive rates. This is considered an effective solution.

IV. Performance Analysis

The performance analysis is described in this section by using the performance metrics like accuracy, delay, throughput, reliability, redundancy, average message overhead, and code rate.

A. Simulation Analysis

The simulation outcomes are analyzed by using the NS2 simulator. NS2 has turned into the most extensively used open-source network simulator, and one of the most commonly used network simulators. The simulation parameters are represented in the Table I provided here.

TABLE I. Simulation Parameters

Parameters	Values
Number of nodes	100
Topology size	2500 X 1000 m^2
Maximum Packet in Interface Queue	100bits
Data aggregation energy	5nJ/bit/signal
Protocol	AODV
Duration of round	20 s
Initial energy	100 J
Idle Power	0.1 J
Received Power	0.0645 J
Transmit Power	1.5 J
Sleep Power	0.01 J
Duration Time	99.001 ms
Channel Type	Wireless Channel
Radio Propagation Mode	Two Ray Ground
Antenna Model	Omni Antenna
Interface Queue Type	Queue/DropTail/PriQueue
Network Interface Type	Phy/WirelessPhy
MAC Type	Mac/802_11

B. Performance Analysis

This section provides the performance evidence of the proposed system about the performance metrics like packet delivery ratio, throughput, redundancy, code rate, total number of control packets, most reliable received bits, and the data packets that are sent.

1) Throughput: The number of data packets (measured in bits) that are directed over the total period of simulation is referred to as the throughput. It is also well-defined as the number of valuable bits to the per unit time progressed by the network to a definite destination from a certain source without considering the overhead and data packets that are transmitted. The throughput can be stated mathematically as follows:

$$Throughput = \frac{Number\ of\ data\ packets\ sent\ (bits)}{simulation\ Time\ (secs)} \quad (1)$$

The simulation results are analyzed by simulation time vs throughput for both existing and proposed systems in Table II and Fig. 2 which specifies that the proposed system gives more throughput than the existing system.

TABLE II. Simulation Time vs Throughput

Simulation time	Throughput of existing system	Throughput of the proposed system
5	350	379
10	379	394
15	395	413
20	426	452
25	489	502

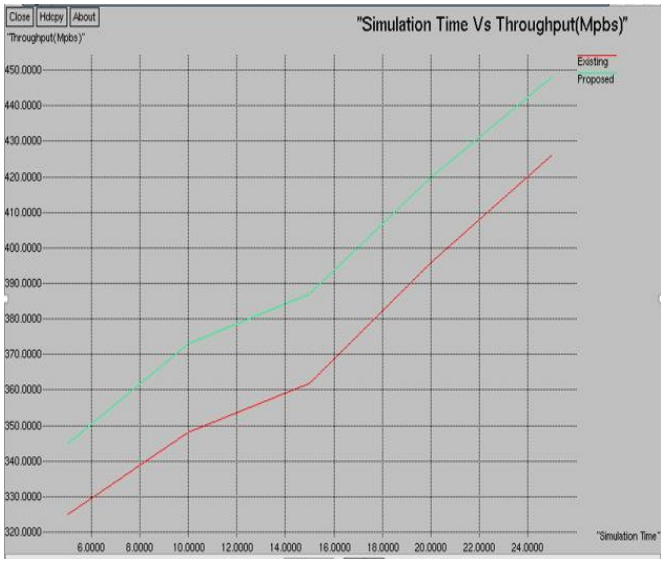


Fig. 2. Simulation Time vs Throughput.

2) Packet Delivery Ratio (PDR): The ratio of the sum of the packets that are received by the destination to the sum of packets that are generated is well-defined as the packet delivery ratio. In other disputes, it is the ratio of the number of data packets received by the receiver to the number of data packets transferred by the source. This can be stated as follows:

$$PDR = \frac{Sum\ of\ packets\ received\ by\ the\ destination}{Sum\ of\ packets\ generated\ in\ the\ source} \quad (2)$$

The simulation results are performed for simulation time vs packet delivery ratio in Table III and Fig. 3 which shows that the proposed system is offering high rate of packet delivery ratio than the existing system.

3) Average end-to-end delay: The average amount of time taken by the packet to reach the client node from the server node is known as the end-to-end delay of the packet. This can be stated as follows:

$$Delay = \frac{Number\ of\ packets\ sent}{simulation\ time} \quad (3)$$

The simulation outcome of the delay is provided for both existing and proposed systems in Table IV and Fig. 4. The delay of the proposed system increases regarding time.

TABLE III. SIMULATION TIME VS PACKET DELIVERY RATIO

Simulation time	Packet Delivery Ratio of the existing system	Packet Delivery Ratio of the proposed system
5	17	20
10	28	39
15	45	57
20	64	78
25	88	95

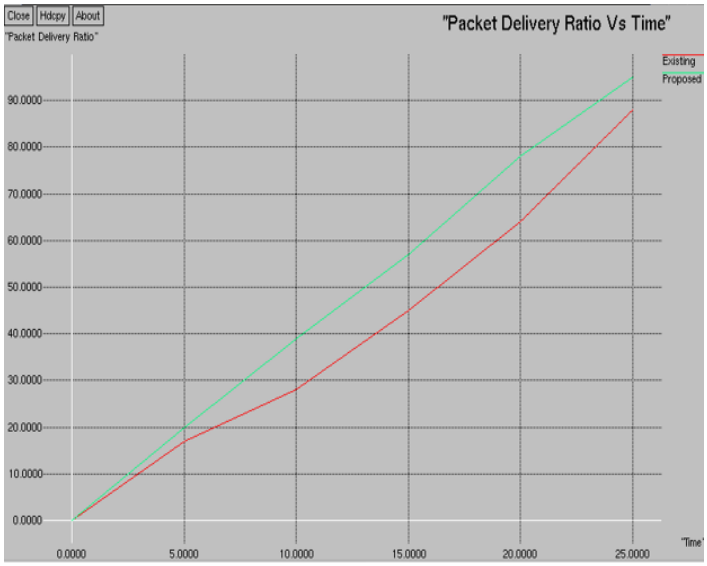


Fig. 3. Simulation Time vs Packet Delivery Ratio.

TABLE IV. SIMULATION TIME VS DELAY

Simulation time	Delay of existing system	Delay of the proposed system
5	73	97
10	174	194
15	245	295
20	366	399
25	468	559

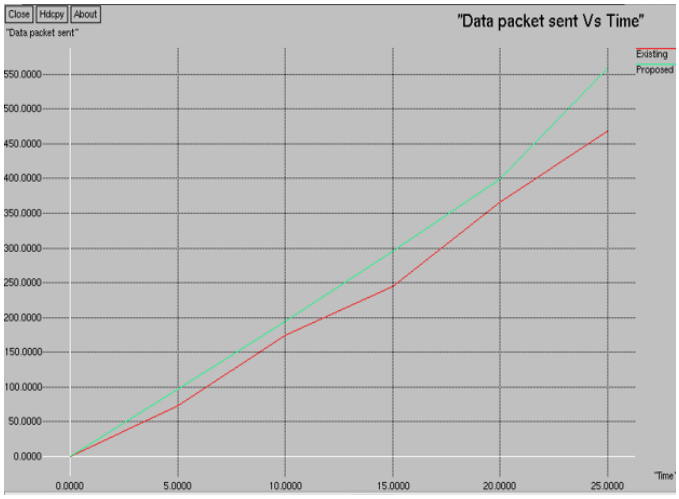


Fig. 4. Average End-to-End Delay vs Simulation Time.

4) Transmission time: The time is taken between the beginnings of transmission till the termination of message or packet transmission is termed as the transmission time. It is the ratio of packet size and the bit rate and can be expressed as follows:

$$Transmission\ time = \frac{packet\ size}{Bit\ rate} \quad (4)$$

The simulation outcome of the code rate is provided for both existing and proposed systems in Table V and Fig. 5. The transmission time of the proposed system increases regarding time.

5) Average message overhead: The ratio of the size of control packets to the total number of data packets that are transmitted successfully to the destination denotes the routing overhead.

By calculating the node and link failures primarily, the route discovery is attained rapidly. Henceforth the message overhead is condensed related to the existing mechanisms Table VI and Fig. 6.

TABLE V. SIMULATION TIME VS CODE RATE

Simulation time	Code rate of existing system	Code rate of the proposed system
4	14	19
8	26	38
12	32	56
16	39	73
20	42	89
24	48	95

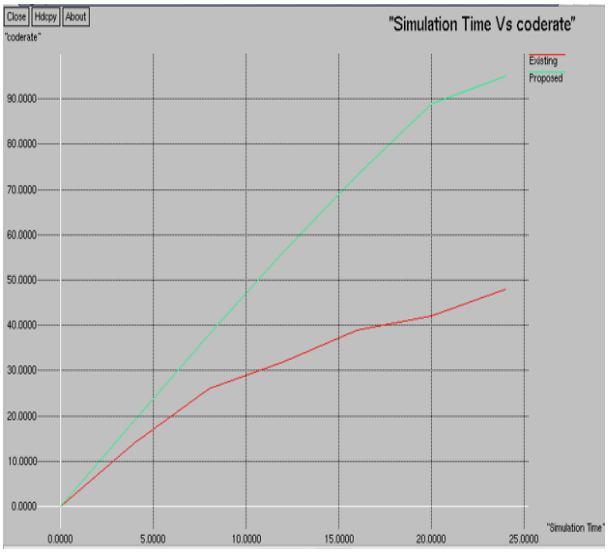


Fig. 5. Simulation Time vs Code Rate.

TABLE VI. SIMULATION TIME VS TOTAL NUMBER OF CONTROL PACKETS

Simulation time	Total number of control packets of existing system	Total number of control packets of the proposed system
4	8	8
8	11	31
12	19	39
16	25	45
20	29	49
24	49	77

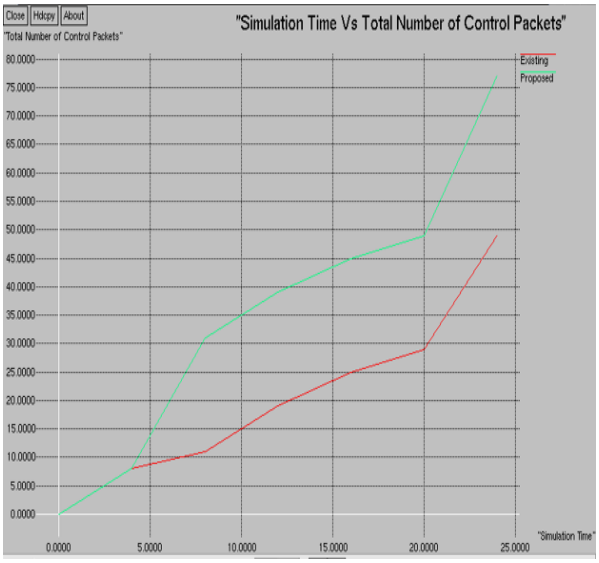


Fig. 6. Simulation Time vs Total Number of Control Packets.

6) Reliability: The ratio of most reliable received bits to the simulation time is identified as the reliability of the system. The simulation outcomes are performed by comparing the reliable bits of both proposed and existing mechanisms which yield a high rate of reliability for the proposed mechanism Table VII and Fig. 7.

TABLE VII. SIMULATION TIME VS RELIABILITY

Simulation time	Reliability of existing system	Reliability of the proposed system
0	0	0
5	65	79
10	159	185
15	229	279
20	348	369
25	427	542

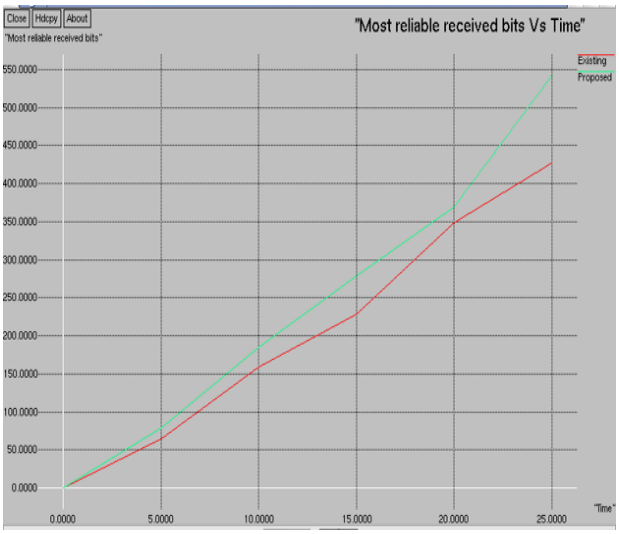


Fig. 7. Simulation Time vs Reliability.

7) Redundancy: Redundancy is well-defined as the rate of occurrence of a fault in the network system which is restrained concerning the time of the simulation.

By this simulation analysis Table VIII and Fig. 8 shows that the existence of faulty nodes is less in the proposed system than the existing mechanism.

TABLE VIII. SIMULATION TIME VS REDUNDANCY

Simulation time	Redundancy of existing system	Redundancy of the proposed system
0	0	0
3	0.255	0.0023
6	0.588	0.0067
9	1.99	0.0133
15	3.65	0.033
20	6.77	0.048
23	9.44	0.069
25	12.54	0.115

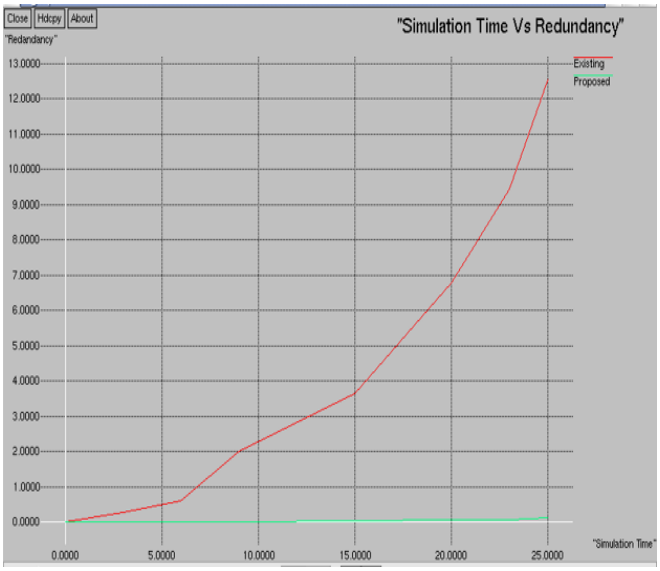


Fig. 8. Simulation Time vs Redundancy.

V. CONCLUSION

MANET was an infrastructure-less fast-growing wireless sensor network in which the routing protocol was complex due to its dynamic nature. The distributed hash table was used to secure the communication of data in which the updating of node status was performed. This in turn causes a storage overhead. The bloom's filter was utilized to compensate the storage overhead of the DHT which was capable of allowing false-positive rates. Hence, the integration of Distributed hash table was done with the bloom filter (DHT-MBF) to compensate for the storage overhead and reduction of false-positive rates. The link stability was considered as an important factor in which the removal of malicious nodes would cause unwanted entry into the other clustering nodes. Therefore, the link stability was enhanced by the modification of Bloom's filter.

The performance analysis was made and the results were obtained by the simulation which shows that our proposed (EDHT-MBF) Enhanced distributed hash table with the integration of modified bloom's filter reduces the storage overhead and occurrence of false positives thereby increasing the security of data communication.

REFERENCES

[1] N. Jain and Y. Chaba, "Simulation based performance analysis of zone routing protocol in Manet," International Journal of Computer Applications, vol. 88, 2014.

[2] C. Cameron, I. Khalil, and Z. Tari, "An ID-based approach to the caching and distribution of peer-to-peer, proxy-based video content," Journal of Network and Computer Applications, vol. 37, pp. 293-314, 2014.

[3] Y. Zhang, D. Li, Z. Sun, F. Zhao, J. Su, and X. Lu, "CSR: Classified Source Routing in DHT-Based Networks," IEEE Transactions on Cloud Computing, 2015.

[4] L. Carrea, A. Vernitski, and M. Reed, "Optimized hash for network path encoding with minimized false positives," Computer networks, vol. 58, pp. 180-191, 2014.

[5] A. Hinds, M. Ngulube, S. Zhu, and H. Al-Aqrabi, "A review of routing protocols for mobile ad-hoc networks (manet)," International journal of information and education technology, vol. 3, p. 1, 2013.

[6] G. Fersi, W. Louati, and M. B. Jemaa, "Distributed Hash table-based routing and data management in wireless sensor networks: a survey," Wireless networks, vol. 19, pp. 219-236, 2013.

[7] S. A. Abid, M. Othman, and N. Shah, "3D P2P overlay over MANETs," Computer Networks, vol. 64, pp. 89-111, 2014.

[8] S. A. Abid, M. Othman, and N. Shah, "Exploiting 3D structure for scalable routing in MANETs," IEEE Communications Letters, vol. 17, pp. 2056-2059, 2013.

[9] W. Kenny and S. Weber, "Below cross-layer: an alternative approach to service discovery for MANETs," in International Conference on Ad Hoc Networks, 2012, pp. 212-225.

[10] Y. Kanizo, D. Hay, and I. Keslassy, "Maximizing the throughput of hash tables in network devices with combined SRAM/DRAM memory," IEEE Transactions on Parallel and Distributed Systems, vol. 26, pp. 796-809, 2015.

[11] A. Moussaoui, F. Semchedine, and A. Boukerram, "A link-state QoS routing protocol based on link stability for Mobile Ad hoc Networks," Journal of Network and Computer Applications, vol. 39, pp. 117-125, 2014.

[12] R. Chen, Y. Li, R. Mitchell, and D.-C. Wang, "Scalable and efficient dual-region based mobility management for ad hoc networks," Ad Hoc Networks, vol. 23, pp. 52-64, 2014.

[13] D. S. Dhenakaran and A. Parvathavarthini, "An overview of routing protocols in mobile ad-hoc network," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.

[14] N. C. Fernandes, M. D. D. Moreira, and O. C. M. B. Duarte, "An efficient and robust addressing protocol for node autoconfiguration in ad hoc networks," IEEE/ACM Transactions on Networking (TON), vol. 21, pp. 845-856, 2013.

[15] M. Rajesh and J. Gnanasekar, "Congestion Control Scheme for Heterogeneous Wireless Ad Hoc Networks Using Self-Adjust Hybrid Model," International Journal of Pure and Applied Mathematics, vol. 116, pp. 537-547, 2017.

[16] Y. T. Yu, M. Gerla, and M. Sanadidi, "Scalable VANET content routing using hierarchical bloom filters," Wireless Communications and Mobile Computing, vol. 15, pp. 1001-1014, 2015.

[17] W. Znaidi, M. Minier, and S. Ubéda, "Hierarchical node replication attacks detection in wireless sensor networks," International Journal of Distributed Sensor Networks, vol. 9, p. 745069, 2013.

[18] M. M. Cisse, N. Usunier, T. Artieres, and P. Gallinari, "Robust bloom filters for large multilabel classification tasks," in Advances in Neural Information Processing Systems, 2013, pp. 1851-1859.

[19] Q. He, Q. Dong, B. Zhao, Y. Wang, and B. Qiang, "P2P Traffic Optimization based on Congestion Distance and DHT," J. Internet Serv. Inf. Secur., vol. 6, pp. 53-69, 2016.

[20] T. Abbas, F. Qamar, M. N. Hindia, R. Hassan, I. Ahmed and M. I. Aslam, "Performance Analysis of Ad Hoc on-Demand Distance Vector Routing Protocol for MANET," 2020 IEEE Student Conference on Research and Development (SCOReD), 2020, pp. 194-199.

[21] U. Srilakshmi, N. Veeraiah, Y. Alotaibi, S. A. Alghamdi, O. I. Khalaf and B. V. Subbayamma, "An Improved Hybrid Secure Multipath Routing Protocol for MANET," in IEEE Access, vol. 9, pp. 163043-163053, 2021.

Spark based Framework for Supervised Classification of Hyperspectral Images

N. Aswini[1]
Division of Computer and Information Sciences
Annamalai University, Annamalainagar, India

R. Ragupathy[2]
Department of Computer Science and Engineering
Annamalai University, Annamalainagar, India

***Abstract*—The advancement of remote sensing sensors acquired large amount of image data easily. Primary aspects of big data, such as volume, velocity, and variety, are represented in the acquired images. Furthermore, standard data processing approaches have different limits when dealing with such large amounts of data. As a result, good machine learning-based algorithms are required to process the data with higher accuracy and lower computational efficiency. Therefore, we propose ANOVA F-test based spectral feature selection method with a distributed implementation of this machine learning algorithm on Spark. Experimental results are obtained using the bench mark datasets acquired using AVIRIS and ROSIS sensors. The performance of Spark MLlib based supervised machine learning techniques are evaluated using the criteria viz., accuracy, specificity, sensitivity, F1-score and execution time. Added to that, we compared the execution time between distributed processing and processing with single processor. The results reveal that the proposed strategy significantly cuts down on analytical time while maintaining classification accuracy.**

***Keywords*—Hyperspectral images; spark; supervised classifiers; spectral features; ANOVA F-test; distributed processing**

I. INTRODUCTION

Advances in recent years, optical sensor technology have provided a wealth of data in terms of achieving necessary spectrum, temporal, and spatial resolutions. Hyperspectral imagery make up a significant portion of the spectral details (HSIs) [1].

The currently available high spectral resolution helps us to obtain small materials and mild objects with confined spectral bands for different applications such as identification, town planning, agriculture, surveillance, and quantification [2]. Though, remote sensing often relies on hyperspectral imagery (acquired from various satellites or from airborne sensors) which allows capturing simultaneously the radiance at several wavelength bands. These wavelength bands are contiguous and their range is predominantly high. Certainly, these data act as a major performer in big remote sensing data and these has at least these traditional 4V's. The volume, the velocity, the veracity and the variety [3].

Let's begin with the letter V, which stands for Volume. The amount of data collected remotely is increasing in terms of hours and minutes. In recent the years, there has been a phenomenal increase in the data consumption that is heading from terabytes to exabytes. Velocity is the second V. It refers to the process of creating, analysing, and interpreting a large amount of remote sensing data in a short amount of time. The third V stands for Value. Multisource, multitemporal, multispectral, or hyperspectral data can be acquired remotely. The term "multisource" refers to the fact that images can be acquired from a variety of sources, including RADAR, LiDAR, optical, and so on. Images having varying resolutions are referred to as multiresolution (spatial or spectral) So, it is difficult to processing remote sensed data not only because of its large volume of data but also it pre-processing, storage and analysis. Various recent literatures, have proposed many frameworks to deal with these problems. Among these frameworks, one of the popular framework is Spark. MapReduce is a feature of Apache Spark, an open-source parallel computing platform. It gives you the flexibility, scalability, and performance you need to meet the problems of big data. Spark is a library that combines two important libraries. SQL is used to query structured data, while MLlib is used to learn about various learning algorithms and statistical methodologies [4]. Of course, MLlib is, Spark's open-source Machine Learning (ML) library, which contains a number of useful training features. It also supports a variety of languages, including Python, R, Java, and Scala, as well as a high-level API that enhances Spark's ecosystem and simplifies the building of machine learning pipelines. [5].

Supervised classification using ML is the important method to extract related information from hyperspectral images. In general, a supervised classifier learns from a training phase that contains hyperspectral data and its corresponding class labels, then generalises to identify class labels for hyperspectral data outside of the training set.

In current research, the intended architecture is composed of three stages viz., Feature extraction, Feature selection using ANOVA F-test and supervised classifier. For better classification accuracy, it is necessary to work with good number of features. So that, we include Feature extraction and Feature Selection (FS). FS is an important strategy for selecting a subset of features from a large number of characteristics and then reducing the high data dimensionality, which results in the greatest classification accuracy. The success of feature selector algorithms is generally measured by comparing the classification techniques with and without selection of features. Feature selection is predominantly used to decrease the dimension of the original feature by eliminating the redundant and irrelevant features, and also to increase the performance and effectiveness of classification. The best features identified are used to satisfy some specified criteria. When compared to using the entire feature subset, classification with feature selection lowers the learning cost by lowering the number of

features used for learning, as well as it removes the unnecessary, noisy, and redundant data, and also ensures the best learning accuracy. In this study, we used one-way ANOVA F-test statistics to measure resemblances for relevant features and to reduce the data dimensions of feature space by finding the necessary features, with the objective of reduced computational complexity or enhancing classification accuracy, maybe both. A comparative study has been carried out between RDD based supervised techniques like Decision tree (DT) [6], Random Forest (RF) [7] and Logistic regression (LR) [8] using the combination of FS with the supervised classifiers on regular mode. The overall performance of hyperspectral imaging classification has been considerably enhanced as a result of the powerful feature representation learned using various ML classification approaches, according to several studies [9].

The purpose of this work is to implement the hyperspectral imagery with feature selection method in distributed environment. In order to test the improvement of the suggested technique, we choose to utilise four distinct widely accessible datasets based on image acquisition and image resolution. The rest of the research paper is arranged as follows: Section II discusses the feature selection based on ANOVA F-test method; Section III addresses experimental designs; Section IV reports the configuration; Section V describes the dataset related to works; Section VI reports the performance metrics; Section VII describes the results and analysis of the experiment and Section VIII draws the conclusion in brief.

II. Feature Selection Based on Anova F-test

ANOVA compares the mean value between the classes and decides whether any mean value vary from each other[10]. By using the F-test value, we can calculate the difference between the mean. The F- test value can be calculated by the following equation.

$$F = \frac{MS_B}{MS_W} \tag{1}$$

Where, MS_B characterize the group variance and defined by the equation (2).

$$\frac{\sum_i n_i(\overline{x_i} - \bar{x})^2}{m-1} \tag{2}$$

MS_w is defied by

$$\frac{\sum_i n_i(\overline{x_{ij}} - x_i)^2}{n-m} \tag{3}$$

Here, n is the number of samples in i^{th} group. x refers the overall mean value and x_i refers the sample value of the mean.

III. Experimental Design

An RDD based supervised hyperspectral classification with complete spectral features is proposed in this section. Generally, Spectral features contain significant information for differentiating the materials obtained on the land. In order to improve the classifiers training speed and prediction results of image classification we used distributed computing framework where its computing data is fed into the Hadoop Distributed File System (HDFS) [11] and stored there. We used noise reduced spectral bands with its corresponding labels as input. This RDD based supervised classification have training and testing phases. Here, 70 percent of the total data was used for training, with the remaining 30 percent for testing the trained model. The hyperspectral imagery, along with its matching ground truth is saved in HDFS as input during the training stage. Following that, feature selection based on ANOVA, selected spectral features and their values are loaded into a supervised classifier using Distributed Spark ML. Spark MLlib provides many API's with supervised classifiers. It uses Spark's strong distributed engine to scale out classification on huge datasets. The classifier was then trained using the supplied ground truth. The learned model is then built. The residual 30% of samples in the testing are used to create the feature vector. Following that, the feature vector collected from the dataset is provided to the predictive model, which is a trained data from the supervised classifier. For each pixel, this prediction model generates the right class label. Fig. 1 shows how these training and testing procedures are carried out. The classifiers' evaluation is conducted using various elements like overall accuracy, specificity, sensitivity, and F1-score founded with the help of confusion matrix, as well as the predictive model's outcome and each classifier's execution duration.

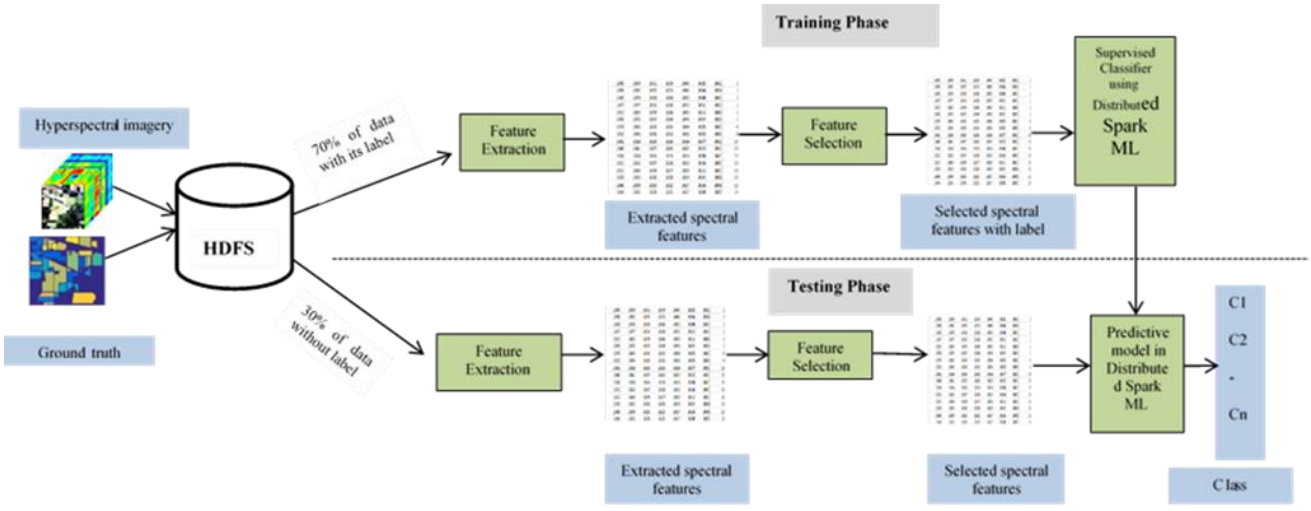


Fig. 1. General Block Diagram of Hyperspectral Image Classification with Feature Selection.

IV. Experimental Configuration

To assess the proposed work described in Section III, we carried out the experiments on single node computer for Hyperspectral image classification focusing on ANOVA based feature selection. The configuration of single node processor intel core i7 7th generation, 16 GB RAM, Apache Spark – 1.6.0, Hadoop 3.2.2 and Linux 18.04.

V. Dataset

Experiments are conducted using commonly available hyperspectral dataset along with its reference ground data are available publicly on the website [12].

A. Indian Pines Dataset

This dataset was gathered throughout a vast region of Indian pines in north-western Indiana. The Airborne visible/infrared imaging spectrometer (AVIRIS) sensor takes images in 224 spectral bands with a spatial resolution of 20nm in the spectral range 0.43 to 0.86nm. Twenty water absorbed bands were removed and remaining 200 were processed for the experiment. Two-thirds of the image is cultivated land, while one-third is woodland. It is 145 x 145 pixels with 16 different categories.

B. Salinas

The second dataset Salinas, was gathered in California's Salinas Valley and has a high spatial resolution. The image, which covers the Salinas area, is 512 x 217 pixels in size and has 224 bands. This scene's ground truth has 16 classes of interest.

C. Salinas-A

Third dataset is Salinas-A, a minimal sub-scene of the Salinas image. It has a resolution of 86 × 83 pixels, 224 bands in the same area, and six classes.

D. University of Pavia

ROSIS sensor absorbed this dataset and it is generated over Pavia, northern Italy. It collects the images in 103 spectral bands with 610 x 610 pixels, and some of the samples contain no information. So, it is removed before analysis. There are nine different types of samples in this scenario ground truth.

VI. Performance Measures

The experimental results of each dataset were assessed using the evaluation metrics such as Accuracy, specificity, sensitivity and F1-score and the result of each classifier is compared with each other [13].

Accuracy: It is the ratio between number of correctly predicted data and total number of input values.

$$\boldsymbol{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (4)$$

Specificity: Specificity refers the proportion between the number of true positive attributes and the number of positive classification results is known as specificity.

$$Specificity = \frac{True\ positives}{True\ positives + False\ positives} \quad (5)$$

Sensitivity: sensitivity is defined by the proportion between the number of true positive results and the total number of relevant samples.

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ neagtives} \quad (6)$$

F1-score: The average of specificity and sensitivity is F1-score. This score varies between [0, 1]. We may determine the number of cases it properly classifies as well as the classifier's robustness from this.

$$F1 - score = 2 * \frac{(sensitivity\ X\ specificity)}{(sensitivity + specificity)} \quad (7)$$

VII. Experimental Analysis

This Section compares the performance classification results obtained using ANOVA feature selection method on different dataset. In order to get the efficiency of ANOVA method[13], classification is carried out using different number of features. Different feature combination was obtained using ANOVA. We let the first trial consist of 5 spectral features; second contains 10; subsequently the trials had number of features 50,100,150,180 and 200 for Indian pines test data. The reason for selecting different sets of features is to ensure that, fewer features could also obtain the comparable classification accuracy.

From Table I, it is evident that the overall classification accuracy is increases with number of features and beyond certain number of features, accuracy is not increasing. Hence, the features with highest accuracy are selected and compared with full features. For Indian Pines dataset we consider 150 number of features. Like that, for Salinas and Salinas-A we considered 150 features and for University of Pavia dataset we considered 100 selected features. The experimentations were carried out on multi-core machine in pseudo-distributed mode to perform classification on a single machine by creating a pseudo cluster (considering each core as a computer node) in spark [14]. To measure the performance, metrics such as execution time, accuracy, specificity, sensitivity and F1-score were computed. The results obtained from Indian pines dataset are tabulated in Table II. The experimentations were carried out on multi-core machine in pseudo-distributed mode to perform classification on a single machine by creating a pseudo cluster (considering each core as a computer node) in spark. To measure the performance, metrics such as execution time, accuracy, specificity, sensitivity and F1-score were computed. The results obtained from Indian pines dataset are tabulated in Table II. From Table II, it can be inferred that RF & DT discloses better accuracy compared with remaining two classifiers. LR and GNB obtains similar accuracy score of 51%. However, LR requires very short execution time than other 3 classifiers. As RF is an ensemble-based method, it produces better results than single classifier DT but RF requires more execution time than DT.

Table III compares the experimental results of Salinas dataset. It is observed that GNB performs poorly and achieved only 32% accuracy. Like Indian pines dataset, DT and RF perform equally good and but LR takes very less execution time.

TABLE I. OVERALL ACCURACY PERFORMANCE OF DIFFERENT NUMBER OF FEATURES ON INDIAN PINES

Supervised classifiers	Number of selected features						
	5 Features	*10 Features*	*50 Features*	*100 Features*	*150 Features*	*180 Features*	*200 Features*
Decision tree	59.16	59.93	59.41	61.69	63.24	61.40	61.40
Random Forest	58.38	59.62	60.68	61.74	64.98	63.69	63.00
Logistic Regression	51.24	51.42	51.05	51.02	51.63	51.65	51.69
Gaussian Naïve Bayes	19.18	21.55	23.74	24.63	30.91	51.41	50.78

TABLE II. EXPERIMENTAL RESULTS OF INDIAN PINES DATASET WITH 180 FEATURES

Metrics	Decision Tree	Random Forest	LR	Gaussian NB
Accuracy	61.40	63.69	51.41	51.41
Sensitivity	67.49	72.98	51.00	47.00
Specificity	67.49	72.98	51.41	47.00
F1-score	67.00	72.90	51.41	47.00
Time	0.0942	0.1111	0.0286	0.1036

TABLE III. EXPERIMENTAL RESULTS OF SALINAS DATASET WITH 180 FEATURES

Metrics	Decision Tree	Random Forest	LR	Gaussian NB
Accuracy	76.98	79.25	51.52	32.00
Sensitivity	62.78	59.51	51.52	30.00
Specificity	62.78	59.51	51.52	30.00
F1-score	62.78	59.51	56.41	47.00
Time	0.1086	0.0837	0.0311	0.0790

Experimental results of Salinas-A dataset are tabulated in Table IV from that we infer that RF achieves 63% of accuracy value and DT scores 61.40% LR and GNB produces equivalent result of 51% LR performs quickly than other classifiers it requires only 0.02 s.

Table V compares the results produced from urban based dataset called Pavia University from that we infer that GNB out performs RF accuracy of GNB depends on feature value but RF executes faster than GNB. LR performs lower than all classifiers. By comparing other metrics like specificity, sensitivity and F1-score value acquired from various classifiers, it is evident that RF performs well than other classifiers.

Fig. 2 illustrates the performance of different classifiers based on their accuracy value using proposed method. As shown in Fig. 2, it is clearly shown that RF performed better than all other classifiers.

Fig. 3 describes the performance of different classifiers based on their specificity value using proposed method. As shown in Fig. 3, it is clearly shown that RF performed better on large datasets like Indian pines, Salinas. DT classifier performed better on smaller dataset like Salinas-A. Traditional classifier GNB performed better than RF on Pavia University dataset.

TABLE IV. EXPERIMENTAL RESULTS OF SALINAS-A DATASET WITH 180 FEATURES

Metrics	Decision Tree	Random Forest	LR	Gaussian NB
Accuracy	61.40	63.69	51.41	51.41
Sensitivity	67.49	72.98	51.00	47.00
Specificity	67.49	72.98	51.41	47.00
F1-score	67.00	72.90	51.41	47.00
Time	0.0984	0.092	0.0225	0.802

TABLE V. EXPERIMENTAL RESULTS OF UNIVERSITY OF PAVIA WITH 100 FEATURES

Metrics	Decision Tree	Random Forest	LR	Gaussian NB
Accuracy	63.00	64.05	43.61	70.04
Sensitivity	62.57	64.00	43.28	70.00
Specificity	62.57	64.00	43.28	69.09
F1-score	62.57	64.00	43.28	69.09
Time	0.0982	0.1003	0.2620	0.7689

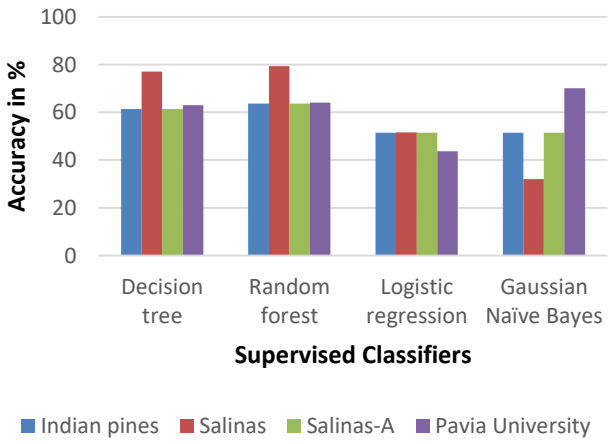


Fig. 2. Performance of Classifiers based on Accuracy Value.

Fig. 4 illustrates the performance of different classifiers based on their sensitivity value using proposed method. As shown in Fig. 4, it is clearly shown that RF performed better than other classifiers on all agricultural data set. GNB performed well on Pavia University dataset.

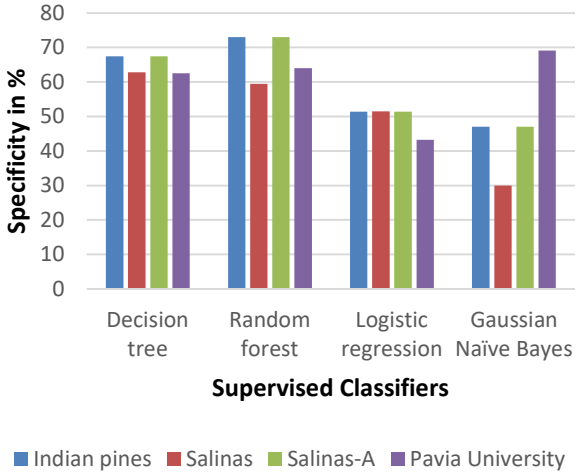


Fig. 3. Performance of Classifiers based on Specificity Value.

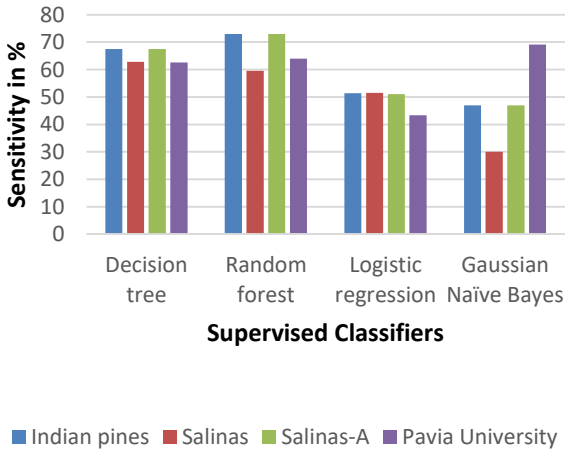


Fig. 4. Performance of Classifiers based on Sensitivity Value.

Fig. 5 illustrates the performance of different classifiers based on their F1-score value using proposed method. As shown in Fig. 5, it is clearly shown that like specificity and sensitivity RF performed better on all agricultural data set. GNB performed well on Pavia University dataset.

Fig. 6 compares the execution time of Indian pines data set on distributed mode using spark MLlib and normal classification method. It clearly shows that, classification using distributed processing reduces the computational time [15].

Fig. 7 compares the execution time of Salinas data set on distributed mode using spark MLlib and normal classification method. As shown in Fig. 7, it clearly shows that, classification using distributed processing reduces the computational time [15].

Fig. 8 compares the execution time of Salinas-A data set on distributed mode using spark MLlib and normal classification method. As shown in Fig. 8, it shows that classification using distributed processing reduces the computational time [15].

Fig. 9 compares the execution time of University of Pavia data set on distributed mode using spark MLlib and normal classification method on different classifiers. As shown in Fig. 9, it clearly shows that, classification using distributed processing reduces the computational time[15].

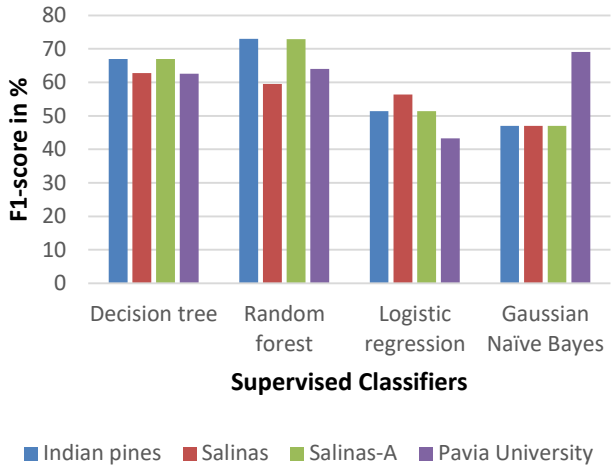


Fig. 5. Performance of Classifiers based on F1-score Value.

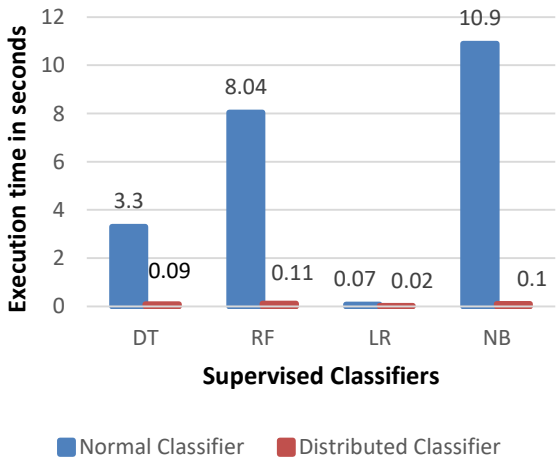


Fig. 6. Execution Time of Various Classifiers on Indian Pines Dataset.

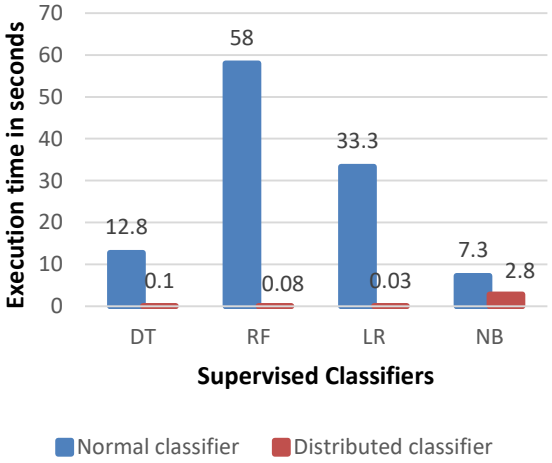


Fig. 7. Execution Time of Various Classifiers on Salinas Dataset.

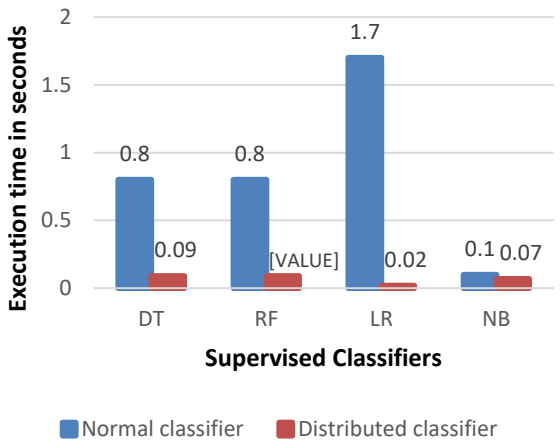


Fig. 8. Execution Time of Various Classifiers on Salinas-A Dataset.

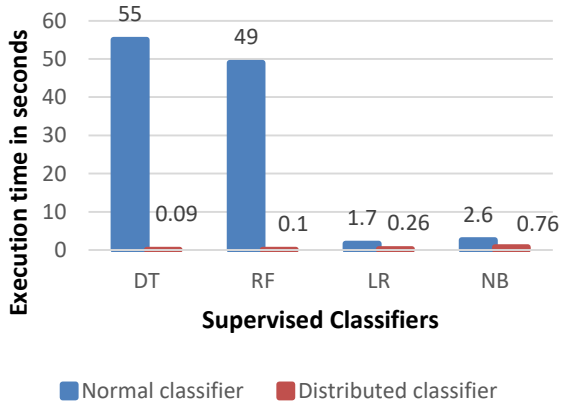


Fig. 9. Execution Time of Various Classifiers on Pavia University Dataset.

VIII. Conclusion

The proposed method uses, Spark based distributed environment for classification of Hyperspectral images with ANOVA feature selection. By comparing it with performance of normal classification methods, the proposed method leads very less computational time and produces good accuracy. Also, we found that distributed method reduces the computational time. As a conclusion remark, Random Forest and Decision tree method of classification produces better accuracy for given hyperspectral dataset. This work uses spectral data for classification of high dimensional hyperspectral image. As a future work, spatial related feature and the fusion of spatial-spectral features can be considered to achieve better classification results with reduced computational time.

References

[1] R. O. Green et al., "Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," Remote Sensing of Environment, vol. 65, no. 3, pp. 227–248, 1998.

[2] R. Ragupathy and N. Aswini, "Performance Comparison of Filter-Based Approaches for Display of High Dynamic Range Hyperspectral Images," in Advances in Intelligent Systems and Computing, 2020, vol. 1079, pp. 79–89.

[3] S. Misra and S. Bera, "Introduction to Big Data Analytics," Smart Grid Technology, pp. 38–48, 2018.

[4] J. A. Richards, "Analysis of remotely sensed data: The formative decades and the future," IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 422–432, 2005.

[5] X. Meng et al., "MLlib: Machine learning in Apache Spark," Journal of Machine Learning Research, vol. 17, pp. 1–7, 2016.

[6] R. Sharma, A. Ghosh, and P. K. Joshi, "Decision tree approach for classification of remotely sensed satellite data using open source support," Journal of Earth System Science, vol. 122, no. 5, pp. 1237–1247, 2013.

[7] S. R. Joelsson, J. A. Benediktsson, and J. R. Sveinsson, "Random Forest Classifiers for Hyperspectral Data," IEEE, pp. 160–163, 2005.

[8] N. Aswini and R. Ragupathy, "On Appraisal of Spectral Features Based Supervised Classifications for Hyperspectral Images," International Journal of Recent Technology and Engineering, vol. 8, no. 6, pp. 593–600, 2020.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification, second edition. 2001.

[10] B. Venkatesh and J. Anuradha, "A review of Feature Selection and its methods," Cybernetics and Information Technologies, vol. 19, no. 1, pp. 3–26, 2019.

[11] "HDFS Architecture Guide." [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[12] "Hyperspectral Remote Sensing Scenes - Grupo de Inteligencia Computacional (GIC)." [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

[13] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way ANOVA F-test for e-mail spam classification," Research Journal of Applied Sciences, Engineering and Technology, vol. 7, no. 3, pp. 625–638, 2014.

[14] "Hadoop - Different Modes of Operation - GeeksforGeeks." [Online]. Available: https://www.geeksforgeeks.org/hadoop-different-modes-of-operation/.

[15] N. Aswini and R. Ragupathy, "ANOVA F-test based Framework for Supervised Classifiers on Classification of Hyperspectral Images," vol. 26, no. 12, pp. 394–403, 2020.

Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus: A Comprehensive Review

Gazi Imtiyaz Ahmad[1]
School of Computer Applications, Lovely Professional University, Punjab, India

Jimmy Singla[2]
School of Computer Science and Engineering, Lovely Professional University, Punjab, India

Anis Ali[3]
Department of Management
College of Business Administration
Prince Sattam Bin Abdulaziz University
Al-Kharj 11942, Saudi Arabia

Aijaz Ahmad Reshi[4]*
Department of Computer Science
College of Computer Science and Engineering
Taibah University
Al-Madinah Al-Munawarah, Saudi Arabia

Anas A. Salameh[5]
Department of Management Information Systems
College of Business Administration
Prince Sattam Bin Abdulaziz University
Al-Kharj 11942
Saudi Arabia

***Abstract*—A comprehensive review of sentiment analysis for code-mixed and switched text corpus of Indian social media using machine learning (ML) approaches, based on recent research studies has been presented in this paper. Code-mixing and switching are linguistic behavior shown by the bilingual/multilingual population, primarily in spoken but also in written communication, especially on social media. Code-mixing involves combining lower linguistic units like words and phrases of a language into the sentences of other language (the base language) and code-switching involves switching to another language, for the length of one sentence or more. In code-mixing and switching, a bilingual person takes one or more words or phrases from one language and introduces them into another language while communicating in that language in spoken or written mode. People nowadays express their views and opinions on several issues on social media. In multilingual countries, people express their views using English as well as their native languages. Several reasons can be attributed to code-mixing. Lack of knowledge in one language on a particular subject, being empathetic, interjection and clarification are some to name. Sentiment analysis of monolingual social media content has been carried out for the last two decades. However, during recent years, Natural Language Processing (NLP) research focus has also shifted towards the exploration of code-mixed data, thereby, making code mixed sentiment analysis an evolving field of research. Systems have been developed using ML techniques to predict the polarity of code-mixed text corpus and to fine tune the existing models to improve their performance.**

Keywords*—Sentiment analysis; code mixing; corpus; deep learning; machine learning; NLP; social media text

I. Introduction

People communicate in their native language or any other natural language having official, national or international status. In a bilingual or multilingual community, people use more than one language simultaneously as their medium of communication. These bilingual people often prefer to use mixed language constructions on the internet and social media platforms to communicate with their friends and relatives informally. The utilization of more than one language in a piece of text, whether through code-mixing or switching (or both), for effective communication, is the hallmark of the social media based text-corpus. With the advent of computers and advancements in technologies people used to analyze and process monolingual text-corpus, using various NLP techniques. NLP is the automatic manipulation of natural language text to decipher useful information. As an area of Artificial Intelligence (AI), NLP deals with training a machine for processing the text for human-computer interaction possible in natural languages [1]. NLP involves the use of computers to process natural language data. The process tends to be just about as straightforward as checking word frequencies to look at changed composing styles or as intricate as understanding total human expressions [2].

Now-a-days, people use social media for various purposes, ranging from daily news and update about the current political and social events, sports, business, entertainment, communicating with family and friends, product/service reviews and opinions and many more [3]. In a bilingual community, people often use more than one language for communicating their perspectives, reviews, opinions and proposals on various subjects. Online Indian language users are exponentially growing and as reported by an investigation by KPMG UK and Google, it is assessed that by 2021, 73% of Indian users would prefer to use native Indian languages. Researchers in the field of NLP have found it quite interesting

*Corresponding Author.

to analyze and decipher information from the text collected from well-known social networking platforms. However, the task is challenging because of a number of reasons. The text present on these platforms is characterized by having spelling errors, Meta tags (hash tags), creative spellings (*f9 for fine*) abbreviations (BTW for by the way) phonetic typing (becoz for because), word plays (*gooood for good*) and so on [4]. All these constraints make it challenging for an NLP researcher to deduce valuable information from the text. Therefore, a considerable percentage of text available on these sites is in languages such as Spanish, Chinese, Arabic, Hindi, Urdu, etc. In the recent past people, especially in bilingual countries like India, not only use a native script to write in their own languages, they also write in the Roman script to express their feelings.

Therefore, people write in code- mixed or code-switched form. Code in communications refers to the rule for converting a piece of information into another form of representation. Code mixing and code- switching are used in bilingual communities where people prefer their native language and a second language in different domains. Although code-switching and code- mixing are usually interchangeable terms in their usage, there are few differences between the two. While code- switching is actually the process of shifting from one language to another, code-mixing on the other hand means the mixing of different phonetic units such as words, phrases, morphemes, clauses, affixes and modifiers of some different language into the expressions of some other language. Thus, the code-switching is inter-sentential, while as code-mixing is intra-sentential which is constrained by grammatical principles.

Example of Code-mixing

Principal appki application ko reject karega. Likh kay leylo.

Translation: The principal will reject your application. Take it from me.

Example of Code-Switching

The principal will reject your application. Likh kay leylo.

Translation: The principal will reject your application. Take it from me.

There are many reasons why people use a multilingual approach while expressing themselves on the web and social media sites. Code mixing and code- switching occurs in informal communication and are used by multilingual speakers. In [5] a list of a number of reasons why code mixing occurs. Bilingualism, speaker and partner speaker, social community, the situation, vocabulary and prestige are the main reasons for code- mixing on social media platforms.

The main reason for code-mixing or switching can be the absence of a specific word or a phrase in a language that necessitates a person to use a word or a phrase from his/her native langue to make the receiver understand it better. The detailed motivation and reasons for code-mixing and code-switching are explained in [5]. On Social media platforms, in a multilingual society, people often mix multiple languages to express their feelings. However, they do not use native language scripts; rather they prefer the roman script to compose non-English words. Automatic language detection in such a scenario is a herculean task. Sentiment Analysis also referred to as opinion mining or emotion analysis, is the identification, recognition or categorization process of people's views and reviews for a service, a product, social issue, an event or a moment into 'positive', 'negative' and 'neutral' classes [6]. Sentiment analysis of dataset containing the data with code-mixed text is a laborious process, ranging from preprocessing of data, language identification to classification. The challenges which need to be addressed before assigning sentiments are posed mainly by unstructured sentences, mixed language constructs, spelling variants, grammatical mistakes, etc0 [7]. Also because of the noisy nature of code-mixed data and the non-availability of annotated resources, sentiment extraction from a code-mixed text has become a challenging task [8]. Therefore, sentiment analysis of the multilingual text has become increasingly an important research area [9]. The general workflow of the Code-Mixed text data Sentiment analysis process is shown in Fig. 1.

A comprehensive review of ML techniques for sentiment analysis of the code-mixed text is presented in this paper. Techniques and approaches of ML and Deep Learning (DL) for bilingual or multilingual text Sentiment Analysis are described along with their corresponding results in different scenarios and using different types of datasets.

The key research highlights of this study are:

- To present the results of recent literature on sentiment analysis of Code-Mixed and Switched languages.
- To provide a systematic review of studies performed on sentiment analysis of Code-Mixed and Switched English with Indian languages.
- To explore and report the current state of research in Code-Mixed and Switched languages using various machine learning and deep learning techniques.
- To present the results of various machine learning models in terms of their performance metrics used by the recent studies in code-Mixed and Switched English with Indian languages.

The paper is organized as follows: Section II and its subsections provide the Machine Learning and deep learning methods used in sentiment analysis of code-mixed social networking data. Section III presents the results of Sentiment Analysis of Code-mixed Indian languages; Section IV presents a discussion of the study. The conclusion is presented in Section V.

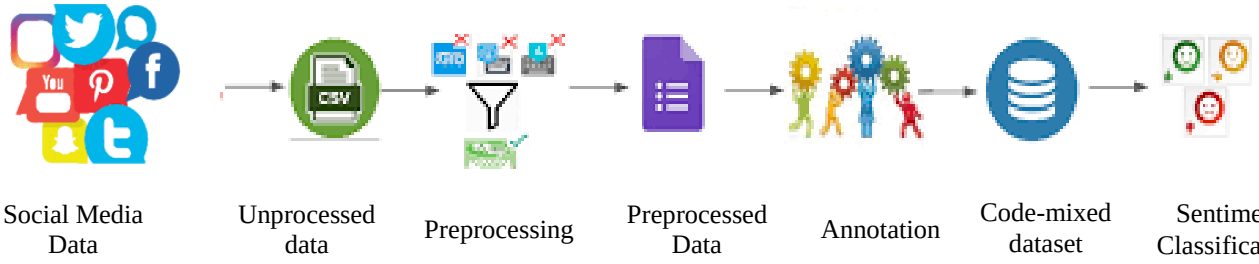


Fig. 1. Sentiment Analysis Process of Code-Mixed Data.

II. Machine Learning and Deep Learning Approaches for Sentiment Analysis

Machine Learning allows computers to seek new tasks without being explicitly programmed to perform them. In Sentiment analysis, ML can be used to analyze text for polarity. Sentiment analysis models have been trained to analyze and understand complex natural language such as human patterns of speech, the context of the sentence, sarcasm, idioms, negation, metaphors, etc. with reasonable and accepted accuracy [10]. Researchers have successfully proposed various approaches for sentiment analysis of English language data using Machine Learning and Deep learning models [11] [12].

Deep Structured Learning commonly known as Deep Learning has acquired a lot of consideration from the recent past in the Machine Learning approach of research [13] Deep learning uses multiple layers to mine higher-level features from the given input data. It is used for a number of applications viz. text analysis, pattern analysis, classification, image processing, etc. and uses non-linear information for feature extraction and transformation in the supervised and unsupervised domain [14]. Deep Learning techniques permit computational models that manage various processing layers to learn representations of data with multiple layers of abstraction. Deep in Deep Learning denotes the layer numbers that form the Neural Network in traditional methods neural networks were of three layers viz. input, output and hidden. The maximum the number of hidden layers, the deep is the neural network [15]. Sentiment Analysis Approaches using Machine Learning and Deep Learning approaches have been illustrated in Fig. 2.

A. Support Vector Machine

Support Vector Machine (SVM), designed by Vladimir Vapnik in 1995 [16], is a non-linear classifier and is a popular and robust classification and regression algorithm for data analysis and pattern [17]. The goal of SVM is to find the best and ideal hyper-plane that maximizes the gap between data points of two unique classes. If the data is un-labeled, Support Vector clustering is used [18]. SVM data classification concept has been illustrated in the plot given in Fig. 3. The support vectors represent the data points which are closest to the hyper-plane with a distance equivalent to margin.

A word-level classification of English-Nepali and English-Spanish code-mixed public network data was proposed in [19]. The authors performed experiments with linear kernel SVM classifier using word and character n-gram features. The model achieved an accuracy of 77.5% for Nepali- English and 80% accuracy for Spanish-English using basic features and applying a 6-way SVM classifier. The authors suggested that the features of Neural Network may improve the accuracy.

A Code mixed Language identification system for social communication text of Tamil-English and Malayalam-English was proposed in [20]. The system identifies the language on the basis of words. By using the character embedding approach, the system used trigram and n-gram features. For training and testing of the model SVM has been used. The proposed model achieved 93% and 95% accuracy for Malayalam-English and Tamil- English data. The authors suggested that availability of more code-mixed data and using trigram features shall be sufficient for the development of a language identification system.

A Hindi-English Sentiment Analysis system for Twitter data to forecast the sentiment present in the data has been proposed in [21]. Researchers have used tf-idf vector and GloVe Vector features along with the Support Vector Regression (SVR) model. The model achieved an f-score of 0.662.

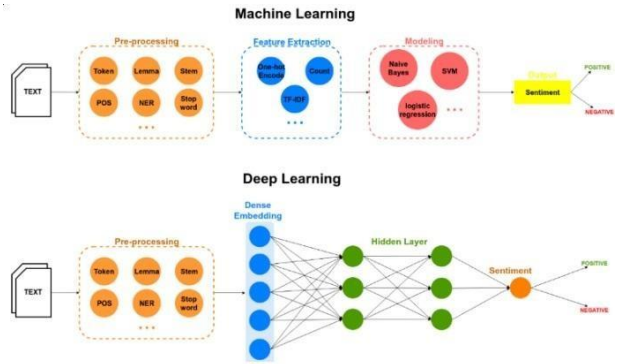


Fig. 2. Sentiment Analysis ML and DL Models.

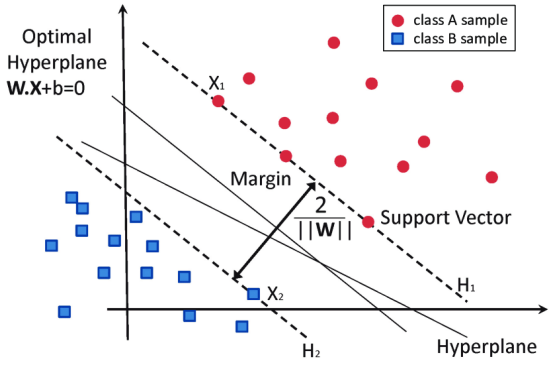


Fig. 3. Classification of Data by SVM.

Shared tasks on Sentiment Analysis of Indian Languages (SAIL) have been organized to identify sentiments in code-mixed datasets collected from media platforms like Twitter, Facebook and other social media platforms of Indian languages, especially language pairs of Hindi-English and Bengali-English [22]. Details of the shared task held during ICON-2017 (the International Conference on Natural Language Processing-2017) were presented by [23]. The goal of the shared task was to identify sentence-level sentiment polarity of code-mixed datasets of language pairs Hindi-English and Bengali-English. The authors presented a detailed overview of problem definition, dataset collection, participant systems and the evaluation process of the shared task. The SVM classifier achieved the best results. Word and character n-grams features were used and applied to SVM classifier for sentiment identification. Thus f-score of 0.569 were achieved for Hi-En and 0.526 for Bi-En datasets.

B. Naïve Bayes

Naïve Bayes (NB), a data mining algorithm [24] is a probabilistic ML classification approach derived from the application of Bayes Theorem with a vast scope in real world applications [25]. The approach assumes that a new object is categorized to a class on the basis of the supposition that all features are independent given in the class [26]. The theorem can be written as in equation 1 and illustrated in Fig. 4.

$$P(A|B) = P(B|A)P(A) \quad (1)$$

Using the probability concept given by Bayesian theorem the equation can be represented as:

$$\text{Posterior} = \frac{\text{Prior x Likelihooh}}{Evidence} \quad (2)$$

NB classifier has been derived from the concept of Bayes Theorem with assumptions of having strong independence between the features.

A system to prepare, collect, filter and identification of sentiment of Twitter data was presented in [27]. The authors applied various supervised ML algorithms viz. Gaussian NB, Bernoulli NB and Multinomial NB for annotation and classification of English-Bengali code-mixed data. The system also applied Code-Mixed Index (CMI), Code-Mixed Factor (CMF) and other language aspects of sentiment classification.

A system to classify Hindi-English and Marathi-English tweets and comments on YouTube using a number of ML algorithms such as NB, SVM and KNN for performance evaluation of each algorithm was designed by [28]. The results reveal that NB and SVM performed better than KNN.

An automatic POS tagging system was proposed by [29]. The authors used coarse-grained and fine-grained social media text collected from Twitter and Facebook for experimentation purposes. Machine learning algorithms such as NB, Conditional Random Forest (CRF), and random forest along with Sequential Minimal Optimization (SMO) were applied for performance comparison. Various features were used in the process which was done on the word context information. The CRF based model thus attained the f1-score of 0.716.

Authors in [30] carried out experiments to construct an English-Punjabi text sentiment classification system. The data was collected from Facebook posts in the agricultural domain. Two classifiers viz. SVM and NB were applied for sentiment identification. Features like unigram and n-gram were applied to the model. The model achieved best accuracy of 85.5% using Naïve Bayes classifier.

A binary sentiment classification model was proposed by [31]. The model used English-Bengali data collected for movie reviews from social networking sites. For the classification and identification of positive and negative sentiments two supervised ML algorithms, NB and SVM were used. The experimental results reveal that if the test and train data are of similar type that is both language data is in Roman script, SVM gives better results. However, overall Naïve Bayes achieved the best accuracy.

C. Decision Tree

Decision tree (DT) is referred to as a non- parametric ML technique of data mining. Decision Tree is commonly used in regression and classification problems such as marketing, sentiment analysis, scientific discovery, fraud detection, etc. [38] is one of the famous supervised ML classification algorithms. The decision tree splits data into two or more sets and important features that create the best split are used and calculated by the algorithm as illustrated in Fig. 5.

An essential part of NLP is POS Tagging. For the English language data, POS tagging is a complex task. However, for code-mixed text data, this is more challenging and is a focused research area in which still needs a significant amount of work to be done for Indian languages code mixed data. An approach for three code-mixed Indian language texts in language pairs (Hindi-English, Hindi-Bengali and Hindi-Telugu) POS tagging was presented by [32]. The authors used ICON-2015 code-mixed data and applied the Decision Tree ML algorithm for code mixed text POS tagging.

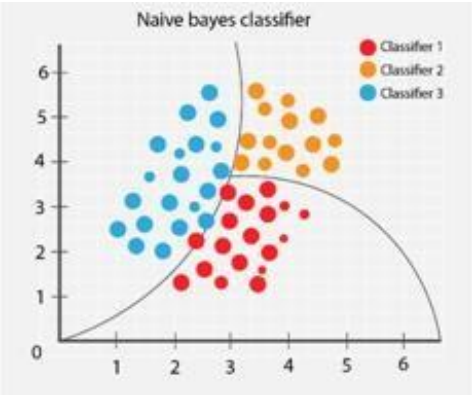


Fig. 4. Naïve Bayes Classification.

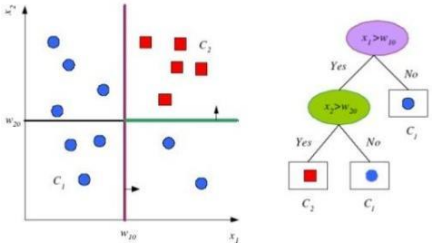


Fig. 5. Decision Tree.

Study on Hinglish (Hindi-English) code-mixed tweets sentiment analysis was done in [33]. Datasets were provided in SemEval-2020 (International Workshop on Semantic Evaluation-2020). The system used J48 Decision Tree as a training classifier and Weka as a tool for the classification. Performance evaluation of the model was done and f-score of 0.53 was achieved.

D. Random Forest

Random forest (RF) is a supervised ML approach used both in for classification and regression problems [34]. It is an ensemble learning algorithm developed by [35]. The algorithm combines DTs and collects their results using averaging. Being a type of supervised learning algorithm, RF has been influenced by [36]. The algorithm works on divide-and-conquer rule. A Random Forest has generally shown excellent performance in scenarios in which the number of observations are less than the number of variables [37]. The general workflow of the technique has been shown in Fig. 6.

For detection of sarcasm, in Hindi-English code- mixed dataset consists of tweets, a baseline supervised classification approach was proposed by [38]. The authors perform 10-fold cross validation using Random Forest classifier. The proposed system also uses Linear SVM classifier and RBF Kernel SVM for the same dataset. However, the RF classifier achieved the better f-score of 78.4.

In collaboration with Forum for Information Retrieval Evaluation (FIRE), a shared task was organized for Code-Mixed Entity Extraction process in Indian Languages (CMEE-IL) in Kolkata, India [39]. Datasets were collected for Hi-En and Ta-En code-mixed social networking data in the said shared task and an Entity Recognition Model was developed. Random Forest Tree Classifier was used for classification. Conditional Random Field Entity Recognition with hybrid features were experimented on the collected corpus. The model achieved 95% of accuracy on training data and a reasonable performance on testing data.

The researchers in [40] have proposed a POS tagger for three Indian code-mixed language pair's viz. Hi-En, Bi-En and Telugu-English. A RF classifier along with a dictionary was applied for fine-grained and coarse-grained datasets consists of tweets, Facebook comments and WhatsApp chats collected from ICON-2016 for the three language pairs. The proposed model achieved best f-score of 78.744 in fine-grained model consisting of Hi-En tweets and 77.944 in coarse-grained model consisting of Bi-En Facebook posts.

E. Artificial Neural Network (ANN)

The concept derived from the human brain in which numerous neurons are interconnected to process data in parallel. ANNs are non-linear mathematical models that show an intricate connection among information sources in order to get a new pattern. ANN can be applied in a range of tasks, including text analysis, image processing, speech recognition, machine interpretation and clinical determination. An ANN has an input layer of neurons or nodes, one or two hidden layers of neurons (or even three), along with a final output layer of neurons. A typical architecture of an ANN is shown in Fig. 7. In a Neural Network the lines connecting nodes (or neurons) are associated with weight. In Fig. 8, a transfer function computes the weighted sum of the inputs while the activation function obtains the result.

The authors of [41] introduced a model for sentiment analysis of Hindi-English text using sub-word level LSTM. The data was collected from Facebook posts and used a 3-class scale of 'positive', 'negative' and 'neutral'. The proposed sub-word level LSTM model achieved higher accuracy than the character-level LSTM model, SVM (Unigram) and Naïve Bayes techniques of machine learning. The overall accuracy of 69.7% was achieved by the proposed system.

Authors in [42] proposed a model in Hindi-English Twitter data for humor detection. Based on models like, Word2Vec and FastText an approach for bilingual word- embedding's applied to BilSTM system for the detection of humor in the text. The proposed approach achieved an accuracy of 73.6%.

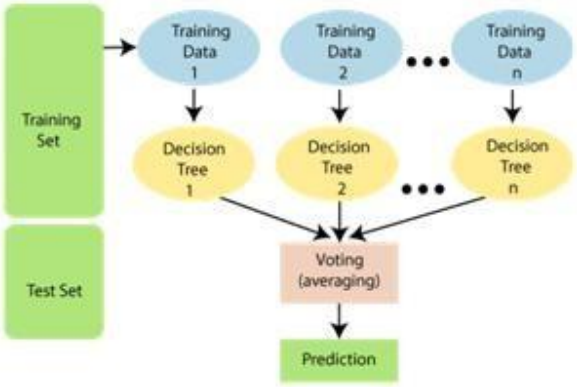


Fig. 6. Random Forest.

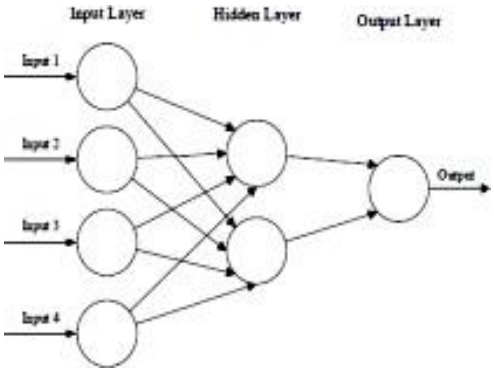


Fig. 7. General Architecture of Artificial Neural.

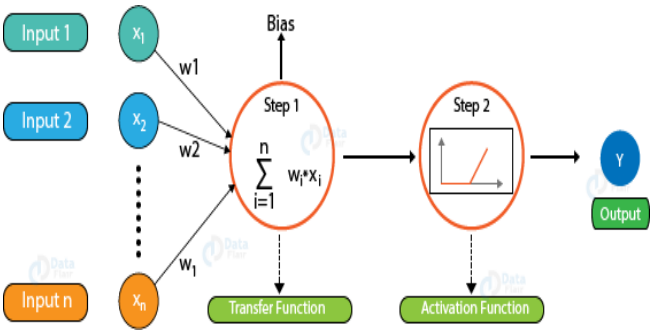


Fig. 8. Transfer Function.

Automatic extraction of sentiments from Hindi- English and Bengali-English Facebook posts was proposed by [43]. The corpus was manually created and annotated. Several preprocessing steps have been employed in order to remove unwanted data from the corpus. A Multilayer Perceptron Model was used for the detection of the sentiment polarity. The proposed model achieved anaccuracy of 68.5%.

F. Convolutional Neural Network

Convolutional Neural Networks (CNN) in recent years have achieved ground-breaking results in a number of pattern recognition fields, ranging from image processing to voice recognition. The most advantageous feature of CNNs is that they reduce the number of parameters in ANNs. This accomplishment has prompted researchers and developers to tackle broader models in order to solve more difficult problems. CNNs are similar to conventional Artificial Neural Networks (ANNs), consisting of neurons that learn to optimize themselves [44]. The neurons obtain inputs to perform operations like the scalar product and non-linear functions, which acts as a foundation for countless Artificial Neural Networks. The complete neural network exhibit a single observant score function from raw input vectors to the final classification output. The general architecture of the CNN for the classification has been illustrated in Fig. 9.

A CNN based system for the sentiment identification of Hindi-English data was proposed by [45]. The sentiment analysis has been done using three class classifications. The classes included 'positive', 'negative' and 'neutral'. The classification of the classes have been done using word-level representations. Since tweets contain informal text, memorization of aspects of the word orthography in a word-level representation was done using CNN. The model achieved an f-score of 0.324 for Hindi- English data.

To compare ML and DL approaches researchers in [46] have used three code-mixed datasets viz. Hindi-English, Bengali- English and Kannada-English. The datasets used in the study have been sourced from Facebook posts and SAIL-2017. A number of Machine and Deep Learning techniques were applied on code-mixing datasets for sentiment analysis. The techniques used include Doc2Vec, SVM, CNN and Bi-LSTM. The experimental results showed that CNN performs better for Kannada-English dataset and achieved an accuracy of 71.5%. The BiLSTM performs better for Hindi-English and Bengali-English datasets with accuracies of 60.22% and 70.20% respectively.

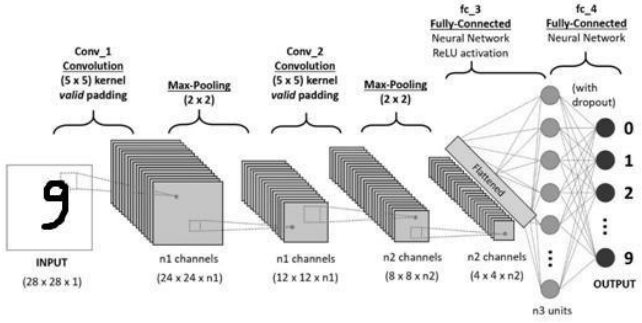


Fig. 9. A Convolutional Neural Network Architecture.

Authors in [47] presented a hybrid model for sentiment analysis of English-Hindi code-mixed data. The method used CNN architecture for generating sub-word level representation for the sentences. Two BiLSTMs, collective encoder and specific encoder are fed with the sub-word level representation. Finally, a Feature Network consists of orthographic features has been combined with the BiLSTMs to achieve an accuracy of 83.5%. The hybrid approach, therefore, combines surface features with Attention-based Recurrent Neural Networks to produce a single representation that can be trained for sentiment classification.

For the identification of emotions in Hindi-English Twitter and Facebook data, authors in [48] proposed a Deep Learning-based system. Several Deep Learning techniques such as 1D-CNN, LSTM, Bi-LSTM, CNN-LSTM and CNN-BilSTM were used to predict the polarity of the sentence. To generate feature vectors, the pre-trained bilingual model was used. The experimental results showed that CNN- BiLSTM model achieved the best accuracy of 83.21%.

The authors of [49] used Facebook comments of Hindi-English code-mixed dataset provided by Trolling, Aggression & Cyber bullying-I (TRAC-I) and apply machine and deep learning models for the classification of text data into a 3-class scale such as, 'Covertly Aggressive', 'Overtly Aggressive' and 'Non-Aggressive' classes. CNN model worked best with an f-score of 0.58 and accuracy of 73.2% as shown in the experimental results.

The study in [50] explores hate speech detection in tweets written in Hindi-English. The authors have used DL models, CNN-ID, LSTM and BiLSTM the semantics detection of hate speech along with the context. The embedding's were generated using Word2Vec. The experimental results were compared with the contemporary approaches. The CNN-ID model outperforms the other two and achieved an overall accuracy of 82.62%.

G. Recurrent Neural Network (RNN)

RNNs are being used by researchers since 1990s. RNN is a neural network with feedback connections is known as a recurrent net [51]. RNN is a form of ANN that works with time series or sequential data. Techniques based on RNNs have been used to solve a broad range of problems. Machine Translation, Speech Recognition, Video Tagging, Text Analysis and Image processing are some examples where RNN algorithms are used. The general architecture of an RNN has been shown in Fig. 10. Each hidden state has hidden nodes also called hidden units.

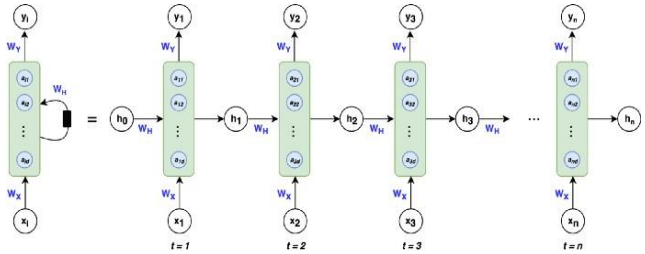


Fig. 10. The General Architecture of an RNN.

An automatic sentiment prediction system of Hindi-English code-mixed dataset consists of tweets was proposed by [52]. The model used the Recurrent Convolutional Neural Network approach to capture the semantics of the text and classify them into three scale classification. The dataset was collected from SemEval- 2020 shared task. An f1-score of 0.69 was achieved by the proposed approach.

The authors in [53] [54] proposed a part-of-speech tagger for Hindi-English, Bengali-English and Telugu-English datasets. The datasets were collected from social networking platforms such as Facebook, Twitter and WhatsApp. The proposed model used Recurrent Neural Network (RNN) to predict word-level part-of-speech tags.

A Sentiment Analysis model of Hindi-English data, based on RNNs, was proposed by [55][56]. Public Facebook pages of popular personalities of Indian Politics and Cinema were used to collect data. The model combines two different BiLSTMs for the identification of sentiment at the sub-word level as well as at the sentence level. The proposed approach used orthogonal features achieved accuracy of 83.5% and f1-score of 0.827.

III. Results of Code-mixed Text Sentiment Analysis for Indian Languages

On Social media sites, netizens in India often use English and their native language such as Hindi in a mixed form to express their opinions on a wide range of topics. Over the years, researchers in the field of NLP have shown keen interest in this new form of text which is often informal and challenging. However, with the advent of NLP tools and techniques, the research related to the analysis of code- mixed textual data has also gained momentum. The significant research studies with their description and the results given for code-mixed text data analysis and sentiment analysis in Indian Languages is given in Table I.

TABLE I. Text Analysis and Sentiment Classification of Code Mixed Text of Indian Languages Gathered from Social Media

S#	Paper/Study	Language	Objective(s)	Dataset(s)	ML/DL Approach	Performance Evaluation
01	Patra, Braja Gopal, et al [25]	Hindi-English Bengali-English	Sentiment Analysis	Tweets	SVM	Language: F1-Score; Hi-En: 0.569; Bi-En: 0.526
02	Ansari & Govilkar [30]	Hindi-English Marathi-English	Sentiment Analysis	Tweets Facebook posts YouTube comments	NB SVM	Model: Language: F1-Score; NB: Hi-En: 0.60; NB: Ma-En: 0.46; SVM: Hi-En: 0.60; SVM: Ma-En: 0.59
03	Jamatia, Anupam et. al. [31]	Hindi-English	Development of annotated corpus, POS tagging, Sentiment Analysis	Tweets, Facebook posts	CRF SMO NB RF	Model: F1-Score; CRF: 0.716; SMO: 0.397; NB: 0.458; RF: 0.706
04	Singh, M et. al.[32]	English-Punjabi	Sentiment Analysis	Tweets, Facebook posts, YouTube comments	NB SVM	Model: Accuracy; NB: 85.5%; SVM: 85%
05	Mandal & Das [33]	English-Bengali	Sentiment Analysis	Movie reviews	NB LR SVM	Model: Accuracy; NB: 59%; LR: 55%; SVM: 57%
06	*Pimpale & Patel [35]*	Hindi-English Hindi-Telugu	POS tagging, Sentiment Analysis	Tweets Facebook posts	NB DT RF	F1-measure; Approach: Hi-Eng: Tal-Eng; NB: 40.4: 46.3; DT: 44.6: 50.3; RF: 43.0: 47.0
07	Ghosh et al [46]	Hindi-English Bengali-English	Sentiment Analysis	Facebook posts	MP	Accuracy: 68.5%

<table>
<tr><td>08</td><td>Sasidhar, T. T et. al.[51]</td><td>Hindi-English</td><td>Development of annotated dataset, classification of emotions</td><td>Tweets
Facebook posts
Instagram comments</td><td>CNN-BiLSTM</td><td>Accuracy: 83.21%</td></tr>
<tr><td>09</td><td>Kumar & Dhar [57]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Facebook posts</td><td>BiLSTM</td><td>Accuracy: 83.54%
F1-score: 0.827.</td></tr>
<tr><td>10</td><td>Baroi, Subhra Jyoti, et al [58]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>Ensemble
LSTM
LSTM + Convolution Layer
BiLSTM
CNN</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>LSTM</td><td>0.5640</td></tr><tr><td>LSTM+Conv</td><td>0.5747</td></tr><tr><td>BiLSTM</td><td>0.576</td></tr><tr><td>CNN</td><td>0.5737</td></tr></table></td></tr>
<tr><td>11</td><td>Veena et. al. [59]</td><td>Hindi-English</td><td>Language Identification</td><td>Facebook posts,
Tweets,
WhatsApp chats</td><td>SVM</td><td>Facebook data (f-score =98.70)
Tweeter data (f-score=93.94)
WhatsApp data (f-score=77.60)</td></tr>
<tr><td>12</td><td>Si, Shukrity, et al. [60]</td><td>Hindi-English</td><td>Aggression detection</td><td>Tweets</td><td>SVM
GBM
LR
Adaboost
DT
KNN
LSTM</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>SVM</td><td>0.5349</td></tr><tr><td>GBM</td><td>0.5410</td></tr><tr><td>LR</td><td>0.5045</td></tr><tr><td>Adaboost</td><td>0.5030</td></tr><tr><td>DT</td><td>0.4938</td></tr><tr><td>KNN</td><td>0.4316</td></tr><tr><td>LSTM</td><td>0.4039</td></tr></table></td></tr>
<tr><td>13</td><td>Soman, K. P. [61]</td><td>Hindi-English
Bengali-English
Telugu-English</td><td>POS) tagging</td><td>Tweets,
Facebook posts</td><td>SVM</td><td><table><tr><td>Language</td><td>Accuracy</td></tr><tr><td>Hi-En</td><td>81.57%</td></tr><tr><td>Bi-En</td><td>76.18%</td></tr><tr><td>Te-En</td><td>68.85%</td></tr></table></td></tr>
<tr><td>14</td><td>Lakshmi & Shambhavi [62].</td><td>English-Kannada</td><td>Word level Language Identification</td><td>Twitter
Facebook posts</td><td>MNB
BNB
SVM
RF
LR</td><td><table><tr><td>Model</td><td>Accuracy</td></tr><tr><td>MNB</td><td>85%</td></tr><tr><td>BNB</td><td>80%</td></tr><tr><td>SVM</td><td>87%</td></tr><tr><td>RF</td><td>78%</td></tr><tr><td>LR</td><td>80%</td></tr></table></td></tr>
<tr><td>15</td><td>Vijay Deepanshu, et al. [63]</td><td>Hindi-English</td><td>Emotion classification</td><td>Tweets</td><td>SVM</td><td>Accuracy : 58.2%</td></tr>
<tr><td>16</td><td>Pravalika et al.[64].</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Movies posts on Facebook</td><td>NB
SVM
DT
RF
MP</td><td>Precision & Recall<table><tr><td>Classifier</td><td>Precision</td><td>Recall</td></tr><tr><td>NB</td><td>0.725</td><td>0.735</td></tr><tr><td>SVM</td><td>0.718</td><td>0.77</td></tr><tr><td>DT</td><td>0.701</td><td>0.723</td></tr><tr><td>RF</td><td>0.752</td><td>0.755</td></tr><tr><td>MP</td><td>0.695</td><td>0.702</td></tr></table></td></tr>
<tr><td>17</td><td>1) Vijay, Deepanshu, et al [65]</td><td>Hindi-English</td><td>Sarcasm detection</td><td>Tweets</td><td>SVM
RF</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>SVM</td><td>0.77</td></tr><tr><td>RF</td><td>0.72</td></tr></table></td></tr>
<tr><td>18</td><td>2) Wu, Wang & Huang [66]</td><td>Hindi-English
Spanish-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>BiLSTM</td><td>F1-score: 0.730</td></tr>
<tr><td>19</td><td>Bhange & Kasliwal [67]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>Ensemble (NB-SVM)</td><td>Accuracy: 0.667
F1-score :0.673</td></tr>
</table>

20	Sharma, & Motlani, [68]	Hindi-English Bengali-English Tamil-English	POS Tagging	Tweets	CRF	Language: Accuracy Hi-En: 80.68% Bi-En: 79.84% Ta-En: 75.48%
21	Sarkar, K [69]	Hindi-English Bengali-English Tamil-English	POS Tagging	Text from social networking sites	Hidden Markov Model	Accuracy : 75.60%
22	*3) Choudhary, Nurendra, et al. [70]*	Hindi-English	Sentiment Analysis	Tweets	siamese network with twin character level Bi-LSTM networks	Accuracy: 78% F1-score: 0.767
23	Singh, K. et. al. [71]	Hindi-English	Language Identification, Entity Recognition	Tweets	CRF LSTM	Model: F1-Score LSTM: 0.693 CRF: 0.767
24	Jamatia, Anupam, et al. [72]	Hindi-English Bengali English	Sentiment Analysis	Tweets	BiLSTM CNN Double BiLSTM Attention Based Model	Language: Model: F1-Score Hi-En: BiLSTM 0.566; D-BiLSTM 0.595; ABM 0.604 Bi-En: BiLSTM 0.623; D-BiLSTM 0.659; ABM 0.675
25	Raha, Tathagata, et al. [73]	Bengali-English	POS Tagging	Tweets	LSTM	Accuracy: 75.29%
26	Parikh, A et. Al[74]	Hindi-English	Sentiment Analysis	Tweets	Ensemble Model (LR,RF, BERT)	F1-score : 0.693.
27	Pratapa, A et. al.[75]	Hindi-English	POS Tagging, Sentiment Analysis	Tweets	LSTM	F1-score : 0.56
28	*4) Kumar, Vaibhav, et al. [76]*	Hindi-English	Language modelling	Social media blogs Facebook Comments	LSTM	Accuracy: 58.9%
29	Bohra, Aditya, et al [77]	Hindi-English	Dataset Creation, Hate speech detection	Tweets	RF	Accuracy: 69.9%
30	Prabhu, Ameya, et al .[78]	Hindi-English	Corpus creation, Sentiment Analysis	Facebook posts	LSTM	Accuracy 69.7%
31	Dahiya, Anirudh, et al. [79]	Hindi-English	Language Identification, POS Tagging, Sentiment Analysis	Facebook Posts	BiLSTM	Accuracy 72.51%
32	Gopal & Das [80]	Hindi-English	Sentiment Analysis	Facebook posts	Ensemble (LSTM MNB)	Accuracy 70.8 F1-Score 0.661

33	Singh & Lefever [81]	Hindi-English	Sentiment Analysis	Tweets	BiLSTM Transfer Learning	Model: LSTM, F1-Score: 0.616; Model: TL, F1-Score: 0.556
34	Santosh & Aravind, [82]	Hindi-English	Hate speech detection	Tweets	SVM RF LSTM	Model: SVM, Accuracy: 70.7%, F1-Score: 0.429; Model: RF, Accuracy: 65.1%, F1-Score: 0.292; Model: LSTM, Accuracy: 66.6%, F1-Score: 0.487
35	Sreelakshmi et. al. [83]	Hindi-English	Hate speech detection	Facebook Posts	SVM RF	Model: SVM, Accuracy: 63.75%; Model: RF, Accuracy: 64.15%

IV. Discussion

Social networking has emerged as an essential part of our lives. It has not only become a platform for individuals to communicate with each other, it also acts as a news media, a platform to connect with people and develop a relationship. It gives an individual the opportunity to express their views on a particular product, service, social movement, government policy etc. Social media thus helps in business and governance tasks. India being the second-largest populous country of the world has a wide range of linguistic diversity. People often express their views in English as well as in their native language resulting in the proliferation of code-mixed data. Mixing of languages or language varieties either in oral or in written form is known as code-mixing.

Sentiment evaluation of social media data analysis plays a crucial role in modern commerce and governance. Classical sentiment analysis systems were developed for dealing with product reviews. With the advancement in NLP tools and technologies, sentiment analysis systems were developed for other tasks as well. Code-mixed text data sentiment analysis is a relatively challenging task right from data gathering to classification. Various studies have been accomplished on "Cross-Lingual Information Retrieval" (CLIR), "Multilingual Information Retrieval" (MLIR) and "Mixed Script Information Retrieval" (MSIR) [84]. In CLIR, a user queries in one language and retrieves desired information in more than one language. In MLIR, a person can query in one or more languages and retrieve information in more than one language. However, the task of retrieval becomes more difficult when dealing with MSIR, due to Romanized text of non-English languages. Also, the social media text contains many non-standard forms such as misspellings, improper use of grammar, letter substitutions, non-standard abbreviations and other ambiguities which makes preprocessing a necessary step in the code-mixed scenario. Various tools for POS tagging, language identification as well as named entity recognition (NER) have been developed for the analysis of code-mixed data over the recent years. However, due to limited datasets particularly annotated datasets for some language pairs and the non-availability of these resources for majority of native Indian languages, and the linguistic catalogues for informal code-mixed text, the automatic text analysis tool development is challenging.

Code-mixed text data analysis in multilingual societies like India has become a vital linguistic research area more specifically for social media content. However, processing such type of data for linguistic analysis is a challenging task due to inherited linguistic complexity and the presence of spelling and grammar variations [85] Therefore, to promote research in code-mixed text, MSIR workshops were organized at FIRE since 2008 [86] various workshops have been conducted on linguistic code-switching computational procedures for language identification and NER textual data for in code- mixing scenarios [87]. SemEval workshops (International Workshop on Semantic Evaluation) have also been conducted. SemEval-2020 was aimed to encourage research in code-mixed Sentiment Analysis of Twitter data.

This paper provides the results of a review study for the sentiment classification of code-mixed Indian languages. The adopted languages, ML/DL/ANN approaches, data sets and challenges in sentiment analysis of code-mixed text data have been highlighted. The results also show that various studies have been carried out in different application domains, thus each of the domains requires different analysis approaches to achieve better performance.

The results show that the most used ML classifier for the sentiment classification of code-mixed Indian language text is SVM followed by NB and RF. Ensemble approaches are also used to classify the code-mixed text. The study also showed that in terms of accuracy and f1- measure, Neural Network approaches perform better than the traditional models. Typically LSTM and BiLSTM algorithms are being used by the researchers for the classification of sentiment in code-mixed datasets. The study reveals that Twitter is the first choice of data collection followed by Facebook and movie/product reviews. Also, appreciable research has been carried out in the Hindi-English public networking site's text followed by Bengali-English. Research has also been carried out in other code-mixed Indian languages such as Punjabi- English, Marathi-English, Telugu-English and Malayalam- English. However, limited or no annotated datasets, text analysis tools and SentiWordNets are not available in most of the code-mixed Indian language text.

V. Conclusion

A comprehensive study of Machine Learning techniques for code-mixed Indian language text collected from popular

media platforms has been carried out in this paper. Among traditional Machine learning approaches, SVM is the first choice of most researchers. In the case of Deep Learning approaches, BiLSTM dominates the research. Twitter data is used for most of the systems and code-mixed social media text for Hindi-English is most researched. Annotated datasets, text and language analysis tools and other lexical recourses are trivial while dealing with code-mixed datasets. In our future work we are going to present a statistical review of Machine Learning approach for Sentiment Analysis of code-mixed social- media text.

REFERENCES

[1] Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. Information Processing & Management, 53(3), 595-607.

[2] Mujahid, M.; Lee, E.; Rustam, F.; Washington, P.B.; Ullah, S.; Reshi, A.A.; Ashraf, I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Appl. Sci.* **2021**, *11*, 8438. https://doi.org/10.3390/app11188438.

[3] Kapoor, K.K., Tamilmani, K., Rana, N.P. et al. Advances in Social Media Research: Past, Present and Future. Inf Syst Front 20, 531–558 (2018). https://doi.org/10.1007/s10796-017-9810-y.

[4] Das, A., & Gambäck, B. (2013). Code-Mixing in Social Media Text. The Last Language Identification Frontier? Trait. Autom. des Langues, 54, 41-64.

[5] Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language, 28(1), 56-75.

[6] Kim, E. (2006). Reasons and motivations for code-mixing and code-switching. Issues in EFL, 4(1), 43-61.

[7] Purba, Y. H., & Suyadi, N. F. (2018). An Anlysis Of Code Mixing On Social Media Networking Used By The Fourth Semester Students Of English Education Study Program Batanghari University In Academic Year 2017/2018. JELT: Journal of English Language Teaching, 2(2), 61-68.

[8] Srivastava, V., & Singh, M. (2020). IIT Gandhinagar at SemEval-2020 task 9: code-mixed sentiment classification using candidate sentence generation and selection. arXiv preprintarXiv:2006.14465.

[9] Kumar, R., & Kaur, J. (2020). Random forest-based sarcastic tweet classification using multiple feature collection. In Multimedia Big Data Computing for IoT Applications (pp. 131- 160). Springer, Singapore.

[10] Nankani, H., Dutta, H., Shrivastava, H., Krishna, P. R., Mahata, D., & Shah, R. R. (2020). Multilingual Sentiment Analysis. In Deep Learning-Based Approaches for Sentiment Analysis (pp. 193-236). Springer, Singapore.

[11] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210.

[12] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631– 1642, 2013.

[13] Reshi AA, Ashraf I, Rustam F, Shahzad HF, Mehmood A, Choi GS. 2021. Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms. PeerJ Computer Science 7:e547.

[14] Furqan Rustam, Aijaz Ahmad Reshi, Wajdi Aljedaani, Abdulaziz Alhossan, Abid Ishaq, Shabana Shafi, Ernesto Lee, Ziyad Alrabiah, Hessa Alsuwailem, Ajaz Ahmad, Vaibhav Rupapara, "Vector mosquito image classification using novel RIFS feature selection and machine learning models for disease epidemiology",Saudi Journal of Biological Sciences, Volume 29, Issue 1,2022,Pages 583-594.

[15] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[16] Aijaz Ahmad Reshi, Furqan Rustam, Arif Mehmood, Abdulaziz Alhossan, Ziyad Alrabiah, Ajaz Ahmad, Hessa Alsuwailem, Gyu Sang Choi, "An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification", Complexity, vol. 2021, Article ID 6621607, 12 pages, 2021.

[17] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[18] A.-Z. Ala'M, A. A. Heidari, M. Habib, H. Faris, I. Aljarah, M. A. Hassonah, Salp chain based optimization of support vector machines and feature weighting for medical diagnostic information systems, in: Evolutionary Machine LearningTechniques, Springer, 2020, pp. 11–34.

[19] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?," ACM SIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 1–13, 2000.

[20] Barman, U., Wagner, J., Chrupała, G., & Foster, J. (2014, October). Dcu-uvt: Word-level language classification with code- mixed data. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 127-132).

[21] Veena, P. V., Kumar, M. A., & Soman, K. P. (2017, September). An effective way of word-level language identification for code-mixed facebook comments using word- embedding via character-embedding. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1552-1556). IEEE.

[22] Garain, A., Mahata, S., & Das, D. (2020, December). JUNLP at SemEval-2020 Task 9: Sentiment analysis of hindi-english code mixed data using grid search cross validation. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1276- 1280).

[23] Patra, B. G., Das, D., Das, A., & Prasath, R. (2015, December). Shared task on sentiment analysis in Indian languages (sail) tweets-an overview. In International Conference on Mining Intelligence and Knowledge Exploration (pp. 650-655). Springer, Cham.

[24] Patra, B. G., Das, D., & Das, A. (2018). Sentiment analysis of code-mixed Indian languages: an overview of SAIL_Code-Mixed Shared Task@ ICON-2017. arXiv preprint arXiv:1803.06745.

[25] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[26] Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 192, 105361.

[27] Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis (Vol. 3, pp. 731-739). New York: Wiley.

[28] Mandal, S., Mahata, S. K., & Das, D. (2018). Preparing Bengali-English code-mixed corpus for sentiment analysis of indian languages. arXiv preprint arXiv:1803.04000.

[29] Ansari, M. A., & Govilkar, S. (2018). Sentiment analysis of mixed code for the transliterated hindi and marathi texts. International Journal on Natural Language Computing (IJNLC) Vol, 7.

[30] Jamatia, A., Gambäck, B., & Das, A. (2015). Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. Association for Computational Linguistics.

[31] Singh, M., Goyal, V., & Raj, S. (2019, November). Sentiment analysis of english-punjabi code mixed social media content for agriculture domain. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 352- 357). IEEE.

[32] Mandal, S., & Das, D. (2018). Analyzing roles of classifiers and code-mixed factors for sentiment identification. arXiv preprint arXiv:1801.02581.

[33] Brijain, M., Patel, R., Kushik, M. R., & Rana, K. (2014). A survey on decision tree algorithm for classification.

[34] Pimpale, P. B., & Patel, R. N. (2016). Experiments with POS tagging code-mixed Indian social media text. arXiv preprint arXiv:1610.09799.

[35] Singh, G. (2020). Decision Tree J48 at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text (Hinglish). arXiv preprint arXiv:2008.11398.

[36] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[37] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[38] Ho, T. K. (1998). The random subspace method for constructing

decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.

[39] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

[40] Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A corpus of english-hindi code-mixed tweets for sarcasm detection. arXiv preprint arXiv:1805.11869.

[41] HB, B. G., Kumar, M. A., & Soman, K. P. (2016). Conditional Random Fields for Code Mixed Entity Recognition. In FIRE (Working Notes) (pp. 309-312).

[42] Bhargava, R., Tadikonda, B. V., & Sharma, Y. (2016, December). BITS_Pilani_Team2@ POS Tagging for Code Mixed Indian Social Media. In International Conference on Natural Language Processing.

[43] Prabhu, A., Joshi, A., Shrivastava, M., & Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. arXiv preprint arXiv:1611.00472.

[44] Sane, S. R., Tripathi, S., Sane, K. R., & Mamidi, R. (2019, June). Deep learning techniques for humor detection in Hindi- English code-mixed tweets. In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 57-61).

[45] Ghosh, S., Ghosh, S., & Das, D. (2017). Sentiment identification in code-mixed social media text. arXiv preprint arXiv:1707.01184.

[46] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[47] Aparaschivei, L., Palihovici, A., & Gifu, D. (2020, December). FII-UAIC at SemEval-2020 Task 9: Sentiment analysis for code-mixed social media text using cnn. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 928-933).

[48] Shalini, K., Ganesh, H. B., Kumar, M. A., & Soman, K. P. (2018, September). Sentiment analysis for code-mixed indian social media text with distributed representation. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1126-1131). IEEE.

[49] Lal, Y. K., Kumar, V., Dhar, M., Shrivastava, M., & Koehn, P. (2019, July). De-mixing sentiment from code-mixed text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 371-377).

[50] Sasidhar, T. T., Premjith, B., & Soman, K. P. (2020). Emotion Detection in Hinglish (Hindi+ English) Code-Mixed Social Media Text. Procedia Computer Science, 171, 1346-1352.

[51] Singh, Vinay, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay, and Manish Shrivastava.(2018) "Aggression detection on social media text using deep neural networks." Proceedings of the 2nd Workshop on Abusive Language Online (ALW2): 43-50.

[52] Kamble, Satyajit, and Aditya Joshi.(2018) "Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models." arXiv preprint arXiv:1811.05145.

[53] Patel, R. N., Pimpale, P. B., & Sasikumar, M. (2016). Recurrent neural network based part-of-speech tagger for code- mixed social media text. arXiv preprint arXiv:1611.04989.

[54] Fausett, L. (1994). Fundamentals of Neural Networks Prentice Hall. Englewood Cliffs, NJ, 7632.

[55] Banerjee, S., Ghannay, S., Rosset, S., Vilnat, A., & Rosso, P. (2020). LIMSI_UPV at SemEval-2020 Task 9: Recurrent Convolutional Neural Network for Code-mixed Sentiment Analysis. arXiv preprint arXiv:2008.13173.

[56] Kumar, V., & Dhar, M. (2018). Looking Beyond the Obvious: Code-Mixed Sentiment Analysis (CMSA).

[57] Baroi, S. J., Singh, N., Das, R., & Singh, T. D. (2020, December). NITS-Hinglish-SentiMix at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text Using an Ensemble Model. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1298-1303).

[58] Veena, P. V., Anand Kumar, M., & Soman, K. P. (2018). Character embedding for language identification in Hindi-English code-mixed social media text. Computación y Sistemas, 22(1), 65-74.

[59] Si, S., Datta, A., Banerjee, S., & Naskar, S. K. (2019, July). Aggression detection on multilingual social media text. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[60] Soman, K. P. AMRITA_CEN@ ICON-2015: Part-of-Speech Tagging on Indian Language Mixed Scripts in Social Media. In 12th International Conference on Natural Language Processing.

[61] Lakshmi, B. S., & Shambhavi, B. R. (2017, December). An automatic language identification system for code-mixed English- Kannada social media text. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) (pp. 1-5). IEEE.

[62] Vijay, D., Bohra, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). Corpus creation and emotion prediction for hindi-english code-mixed social media text. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (pp. 128-135).

[63] Pravalika, A., Oza, V., Meghana, N. P., & Kamath, S. S. (2017, July). Domain-specific sentiment analysis approaches for code-mixed social network data. In 2017 8th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE.

[64] Vijay, D., Bohra, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A Dataset for Detecting Irony in Hindi- English Code- Mixed Social Media Text. EMSASW@ ESWC, 2111, 38-46.

[65] Wu, Q., Wang, P., & Huang, C. (2020). MeisterMorxrc at SemEval-2020 Task 9: Fine-tune bert and multitask learning for sentiment analysis of code-mixed tweets. arXiv preprint arXiv:2101.03028.

[66] Bhange, M., & Kasliwal, N. (2020). HinglishNLP: Fine-tuned Language Models for Hinglish Sentiment Detection. arXiv preprint arXiv:2008.09820.

[67] Sharma, A., & Motlani, R. (2015, December). POS tagging for code-mixed Indian social media text: Systems from iiit-h for icon NLP tools contest. In International Conference On Natural Language Processing.

[68] Sarkar, K. (2016). Part-of-speech tagging for code-mixed Indian social media text at ICON 2015. arXiv preprint arXiv:1601.01195.

[69] Choudhary, N., Singh, R., Bindlish, I., & Shrivastava, M. (2018). Sentiment analysis of code-mixed languages leveraging resource rich languages. arXiv preprint arXiv:1804.00806.

[70] Singh, K., Sen, I., & Kumaraguru, P. (2018, July). Language identification and named entity recognition in hinglish code mixed tweets. In Proceedings of ACL 2018, Student Research Workshop (pp. 52-58).

[71] Jamatia, A., Swamy, S., Gambäck, B., Das, A., & Debbarma, S. (2020). Deep Learning Based Sentiment Analysis in a Code- Mixed English-Hindi and English-Bengali Social Media Corpus. International journal on artificial intelligence tools, 29(5).

[72] Raha, T., Mahata, S. K., Das, D., & Bandyopadhyay, S. (2020). Development of POS tagger for English-Bengali Code- Mixed data. arXiv preprint arXiv:2007.14576.

[73] Parikh, A., Bisht, A. S., & Majumder, P. (2020, December). IRLab_DAIICT at SemEval-2020 Task 9: Machine Learning and Deep Learning Methods for Sentiment Analysis of Code-Mixed Tweets. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1265-1269).

[74] Pratapa, A., Choudhury, M., & Sitaram, S. (2018). Word embeddings for code-mixed language processing. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 3067-3072).

[75] Kumar, V., Pasari, S., Patil, V. P., & Seniaray, S. (2020, July). Machine Learning based Language Modelling of Code Switched Data. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 552-557). IEEE.

[76] Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media (pp. 36-41).

[77] Prabhu, Ameya, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. (2016) "Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text." arXiv preprint arXiv :1611.00472.

[78] Dahiya, A., Battan, N., Shrivastava, M., & Sharma, D. M. (2019, August). Curriculum Learning Strategies for Hindi-English Code-Mixed Sentiment Analysis. In International Joint Conference on Artificial Intelligence (pp. 177-189). Springer, Cham.

[79] Gopal Jhanwar, M., & Das, A. (2018). An Ensemble Model for Sentiment Analysis of Hindi-English Code-Mixed Data. arXiv e-prints, arXiv-1806.

[80] Singh, P., & Lefever, E. (2020, May). Sentiment Analysis for Hinglish Code-mixed Tweets by means of Cross-lingual Word Embedding's. In Proceedings of the The 4th Workshop on Computational Approaches to Code Switching (pp. 45-51).

[81] Santosh, T. Y. S. S., & Aravind, K. V. S. (2019, January). Hate speech detection in Hindi-English code-mixed social media text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 310- 313).

[82] Sreelakshmi, K., Premjith, B., & Soman, K. P. (2020). Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Computer Science, 171, 737-744.

[83] Chakma, K., & Das, A. (2016). Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets. Computación y Sistemas, 20(3), 425-434.

[84] Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014, October). "I am borrowing ya mixing?" An Analysis of English- Hindi Code Mixing in Facebook. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 116-126).

[85] Somnath Banerjee, Kunal Chakma, Sudip Kumar Naskar, Amitava Das, Paolo Rosso, Sivaji Bandyopadhyay, and Monojit Choudhury. 2016. Overview of the mixed script information retrieval (MSIR). In Proceedings of FIRE 2016. FIRE, December.

[86] Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., ... & Fung, P. (2014, October). Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 62-72).

[87] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M.,. & Eryiğit,(2016, January). Semeval-2016 task 5: Aspect based sentiment analysis. In International workshop on semantic evaluation (pp. 19-30).

Hybrid Routing Topology Control for Node Energy Minimization for WSN

K Abdul Basith, T.N. Shankar
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

***Abstract*—Wireless sensing has become an essential feature for minimizing energy for WSN applications. The foundation of the WSN is to implicate the uniqueness of the design feature capabilities, which are tied to different applications choices of interest. The implementation of pervasive algorithms with ubiquitous Network features are depicted with changes in frequency topology bands and congestion regression of the Network. The Network would affect the parametric criteria such as bitrate, Cluster-head energy, minimum energy and bandwidth usage. Our improved hybrid Pervasive algorithms would prevent the different attacks and control with the least tolerant error since topology becomes an integral part of the design, providing efficient Routing for the Network. In order to effectively solve the problem, a hybrid tangential transform with improved topologies for effecting network parameters. The other algorithm implicates the energy-efficient with optimization of stochastic conditional inequity for different network sizes. Performance characteristics of the proposed algorithms for WSN would estimate a tolerant error with a factor of 12% on each feature of the network parameter.**

Keywords*—SCI (stochastic conditional inequality); LEACH; clustering; DDOS (distributed denial of service); DEEC (distributed energy efficient clustering); TETRA (terrestrial trunked radio); WSN (wireless sensor network)

I. Introduction

The Wireless Sensor Network (WSN) affects everyday life as a seed for intelligent applications and ubiquitous systems. WSNs employ traditional WiFi antennas and a collection of internet-connected devices known as "smart nodes" capable of detecting and recording environmental phenomena and physical conditions such as humidity, temperature, pressure, and pollution levels.

Because of the technology's flexibility and communication capabilities, it is possible to exploit the accurate information generated by agent devices. Consequently, the communication is reliable because [1] and [2]. To ensure the long-term survival of innovative applications like IoT (Internet of Things) and military applications, WSNs have been used to demonstrate various intriguing ways. As shown in Fig 1, many smart apps rely on WSNs as their basis. The advent of different scenarios of WSN, which might include the different nodes and that will be deployed with dynamic nodes on each MANETS [3] where ADHOC features are improvised on multi-dimensionality feature [4]. Wireless sensor nodes and a data centre or sink node are the two sensors in this system [5]. It is these nodes that introduce three critical operations into the system.

- Data collection,
- Data processing, and
- Data transmission.

A sensor unit, a processing unit, a data storage unit, a radio transceiver unit, an energy unit, and a power generator are also included as functional modules for controlling and monitoring them. WSN incorporates the sending and reception of data from the data center or sinks node via a wireless channel [6, 9]. With the rise of countries with less established infrastructure, investment in wireless sensor networks (WSN) has become an inescapable consequence because of their low cost and high communication capabilities [10-13]. There are still substantial challenges in WSN related to network capacity restrictions, increasing data loss and collision rates [7]. By offering an efficient routing protocol based on clustering, it is possible to fulfil the Quality of Service (QoS) criteria while at the same time enhancing all-around network performance [8]. Recent years have seen an increased interest in WSN because of its adaptability and communication capabilities, particularly in the absence of conventional networks like Long-Term Evolution (LTE) or Terrestrial Trunked Radio (TETRA), both in academic research and in the IT sector [14-17]. As a result of these concerns and obstacles, the WSN's potential performance may be severely affected by factors such as frequent network topology changes and longer delays in reaching the final destination, route coupling and high packet loss. In order to lower network bandwidth and power consumption while simultaneously extending the lifespan of the Network, new approaches to congestion reduction are needed [18]. Also mentioned in this part are several popular strategies for dealing with congestion control issues and an overview of the topic in network layer congestion management [20].

A. Contribution

This study's primary objective implicates the different scenarios of attacks and its control feature using the pervasive algorithm for improving the overall performance features as throughput, end-end delay time, and the energy minimization for the proposed algorithms. This paper mainly contributes with implementation of routing cluster analysis with proposed route structures from III.B as stated below:

B. Problem Statement

1) Implement the Hybrid Routing protocol for minimum cluster head energy.

2) Improvise a design model on each Routing area calculations for minimum distance measurement.

3) Estimate a stochastic model, for optimizing the error and formulate the relative error features on each node based on SCI algorithm.

4) Our proposed model, with DEEC-SCI and HTTA models have improvised to implicate different performance factors such as bit rate, energy minimization at nodes, and error optimization in dB's.

5) Finally, comparison of LEACH algorithms, TEEN algorithm, and DEEC {SCI-HTTA} (proposed algorithms) have been implicated as Tabulated graphs.

C. Overview

In Section-1, our design implicates different MANET's and their importance of congestion problem and its related features that are governed with Route optimization and region cluster analysis. Network topologies are estimated with different diagram feature consideration and their literature survey of different algorithms with routing protocols list. In section 3, we ensure the different possibilities of optimization algorithms that are governed for ensuring optimal solutions. Finally, we introduce the SCI-HTTA algorithm for energy minimization formulations.

II. Existing Design

A. Concept

Congestion-aware clustering and Routing (CCR) is the purpose of this architecture is to improve network performance by lowering end-to-end delay time, boosting delivery ratio, and prolonging the network lifespan. As a result, several obstacles and concerns must be addressed to fulfil these goals, including the dependency on batteries, the capacity of storage units, and the need to send data to a specific receiver node many times [21].

- Low Overhead: Because the setup phase is done just once in the first round, the overhead associated with executing the setup step in each subsequent round is minimized. It is used during the setup phase to split the network area into levels and sectors, which are subsequently utilized to build clusters of nodes with an equal number of nodes once they have been created.
- Cluster head node (CH) load distribution: As the functions of principal cluster head (PCH) and secondary cluster head (SCH) rotate among all nodes in the Cluster at the start of each cycle, the cluster head node (CH) role is distributed over all nodes, hence reducing the strain on all nodes.
- Stability is achieved when data transfer with SCH-PCH forms Cluster heads correctly.
- Reliability: The values of PCH and SCH should be more reliable and dependable for optimal distance calculations.
- Scalability refers to adding new nodes at any point during a round.
- Fault-tolerance: increasing the packet delivery ratio through the use of fault tolerance solutions is known as fault tolerance.

B. Congestion Problem

When a large number of sensor nodes submit data to a single sink node, there is a high likelihood of network congestion. Several factors contribute to this, including a relatively restricted bandwidth supply and finite network capacity [6]. Figure 2 is an illustration of this phenomenon.

For congestion to arise in a WSN, there are two primary reasons for this: a lack of node capacity and the characteristics of the wireless channel. Due to the prominent restricted resources, sluggish processor, and restricted energy of nodes, congestion in WSN occurs in the first place in the Network's nodes. Secondly, network congestion occurs in WSN because of the nature of the Network, its event-driven nature, channel interference, and the pace at which data is sent and received from the Network. Consequently, protocols developed for WSNs must be lightweight and scalable to maximize the Network's lifespan [22].

C. Cluster-Based Route Optimization Protocols with Block-Based Cluster Formation

The Low-Energy Adaptive Clustering Hierarchy Protocol (LEACH) [14] is a self-organizing, adaptive clustering protocol that uses little energy to perform its tasks. LEACH's procedure is divided into several rounds that are repeated repeatedly. Nodes form themselves into clusters at the end of each round. Each Cluster consists of a single CH node and a large number of MNs nodes, where the CH node receives data from the MNs nodes and executes data tasks simultaneously on the data before sending the aggregated data to the base station. It had several advantages, including balancing energy usage, using TDMA on the MAC, aggregating data from CH nodes, which resulted in a reduction in the high volume of traffic and savings in energy. In addition, it can create new nodes and eliminate dead nodes in each cycle. Nevertheless, it has some disadvantages, such as random selection of CHs, residual energy that is not considered when selecting a CH, single-hop inter-cluster Routing, which increases energy consumption in large-regional networks, and dynamic Cluster which adds extra overhead to the overall network design. Fig 2 is a collection of the most recent successors to the LEACH protocol, together with information on its approach, advantages, and disadvantages.

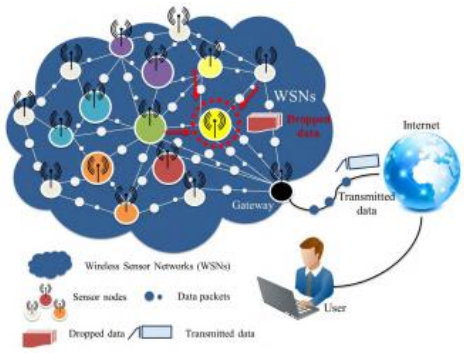


Fig. 1. Representing the Congestion Problem for WSN Systems.

SNO	ALGORITHMS	TECHNIQUE	ADVANTAGES	RESEARCH GAP
1	LEACH-C [14]	Implicate a cental algorithm to form clusters	Improved Performance by 20% to 40%	Every Node requires enables positional analysis from GPS , ence increases extra energy
2	LEACH-F[15]	This design requires a two level of transmission.	No setup overhead	Consumes much power where new nodes and impossible to remove dead nodes.
3	LEACH-E[16]	The design on Cluster head and its selections implicates with amount of resudal energy.	Increased the network lifetime by 40%.	Fixed time round leads to the waste of energy
4	LEACH-ET[17]	The design for this leach model utilzies threshold to increase the proportion stedy state in each round.	Energy efficieny enhanced	This model consumes more energy in transmitting continous message.
5	LEACH-MH[18]	Provides inter and intra multi hop communication model to send and recive data from the devices implanted.	Improvise a large section of energy reduction for large size of network.	Suffers from hot spots and limited scalability

Fig. 2. Representing the Tabulated figure for List of Algorithms for Survey Model.

D. Methodology

With consideration of different design models on the design, LEACH models and its different types have been implicated with Fig-2. According to the current design which effectively improvise on congestion problem and Route area-based design features have been implicated on figure 3 and figure 4. Now, these LEACH models effectively improvise the Routing concept on the basis with region model and its collective intra-cluster routing algorithm mentioned in section F. The current design effectively postulates on different distance formulation with area and its effective scheduling. This feature estimation is represented with on algorithm 2.

In [19], LEACH protocol with fixed number of Cluster head with each round based on random time interval. The P-LEACH is proposed with optimal cluster-based chain protocol which implicates the improvement of PEGASIS and LEACH protocols as hybrid model for optimizing energy. An NS2 feature implementation of the WSN model with P-LEACH is performed and its energy feature reduction is affected on each set of optimal nodes considered [26]. In [27], the authors suggest a hybrid multihop routing protocol that attempts to increase the lifespan of a WSN that is deployed in a globally spread network while maintaining its performance. Despite the fact that its performance has been shown to be superior, the chain-based CH selection and routing of data over MST increases the network's energy consumption, hence shortening its lifespan. This concept creates a hybridization of the metaheuristic cluster-based routing (HMBCR) approach for use in wireless sensor networks. The HMBCR approach begins with a brainstorm optimization with levy distribution (BSO-LD) based clustering procedure that incorporates a fitness function including a fitness function comprising [28].

The event data flow in IMS applications must be delivered in a timely and reliable manner in order for the applications to respond quickly with the relevant actions. However, because of a sensor node's limited energy supply, it is necessary to make a trade-off between latency and energy consumption while selecting the best path to the base station. With the advent of event data traffic in IMS, a multi-objective ant-colony optimization-based quality of service (QoS) aware cross-layer routing (MACO-QCR) protocol has been proposed for inter-cluster communication in WSN-based IMS in order to deal with the multi-constrained routing problem introduced by event data traffic. An improvement to the ACO algorithm is that it is now a multi-objective routing algorithm that considers the energy consumption cost and the end-to-end delay cost of a routing path as two optimization objectives, and in which a routing path is produced through the use of multi-pheromone information and multi-heuristic information that is comprised of two objective functions is used to generate the routing path [29].

In this design, the network itself is optimized for energy efficiency and routing endurance without taking into account the influence of the external environment, resulting in a network that is unable to adapt to environmental changes in a timely way. Thus, the routing survival of various routing protocols in severe conditions is a matter of some debate.

The SMRP (sustained multipath routing protocol), in which routing choices are determined based on a mixed potential field in terms of depth, residual energy, and environment To summarize, the core concept of SMRP is that it instructs messages to pick pathways that make a compromise between delivery delay, energy balance, and routing survival, among other factors [30].

E. Protocols for Region Cluster Analysis and Scheduling

A vertical seam or line of terminals is defined as an Edge D2d Communication (LBDD) [20-25] system, which divides the deployment area into two halves by dividing it vertically. In this context, the nodes positioned on this strip or line are inline nodes, which means that they are located on the strip or line. Information is captured and kept on this line to be accessed later if necessary. It is necessary to convey data from sensors to the line, where it is maintained in the first node that comes into contact with the data. It is possible to send a data query to the line, which will then propagate across the line until it reaches the cluster nodes storing and retrieving data from the inline storage system, at which point the inquiry will be stopped. The data is delivered straight to the sink through the inline node in the next phase, and the multicast routing process is brought to a close. It is believed that each node is aware of its geographic location and the geographic boundaries of the Network. a. Additionally, in addition to the fact that it is relatively simple to identify and configure, LBDD has the benefits of being easily accessible by both sources and sink nodes and having a reasonably low overhead for completing the operations mentioned above. There are some drawbacks to LBDD, such as the fact that it still relies on a live broadcast for spreading metadata along a line and that the line must be wide enough to accommodate hot spots. As a result of this, especially for extensive systems, the flooding on the line will result in a rise in total energy.

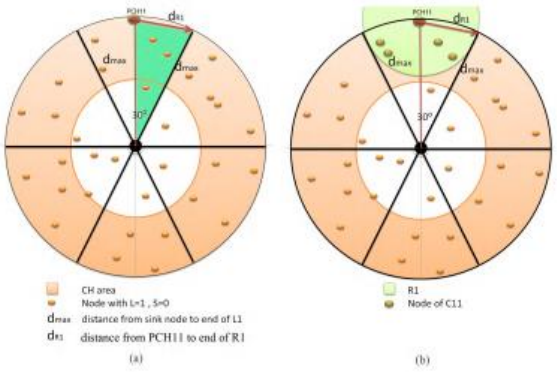


Fig. 3. Representing Area-based Protocol with the Optimum Route (a) Cluster Head Area with Minimum Distance for PCH, (b) Representing the Region of Cluster Nodes in R1 Region.

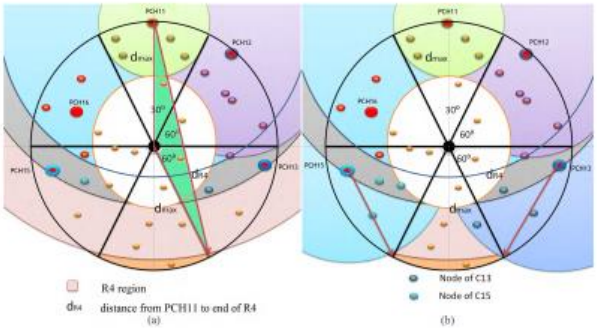


Fig. 4. Representing other Regions R4 with (a) Distance with PCH and other R4 (b) Node with Clusters with C11 and C15.

In Fig 3a&4a, this design implicates each feature of improvements of region clusters with distance optimization on each section of segments with 30 degrees or any other angle ∅. Six sectors of the circular region are estimated with the different clusters and their distances D11 as calculated below.

$$D_{11} = \sum_{i=1}^{N} D_i * (x_i \cos(\emptyset) \pm y_i \sin(\emptyset)) \quad (1)$$

Similarly, Fig 3b-4b considers the distances from clusters and within subdivision region of the Network considered. Hence, the final distance is optimized as:

$$D_x = \sqrt{(D11\cos(\rho) + D11\sin(\rho)} \quad (2)$$

Setup Phase

F. Setup Phase

A single instance of the setup phase is performed in the first round, during which fixed clusters are arranged. In order to start the data transception of WSN, a list of clusters are generated with the capacity considered. Following this phase, the network area is subdivided into levels and sectors. A cluster is formed when each level and sector intersect with one another. It is also necessary for each node to know its cluster number, which is composed of a level number L followed by a sector number S, indicated by the letters Cls. Each Cluster must contain a PCH node and an optional SCH node. The PCH and SCH nodes are picked during the setup phase based on their distance.

Algorithm 1:

Input1: Initiate the number of dead nodes as dn, DC, N

Output: Create a cluster of Nodes for dead

1. Merge the clusters for distance condition not satisfied as
   a. $L(d_c) < lastlevel$
   b. $merge\ level(L(dc), L(dc+1))$
2. Iterate for each l belongs to L
3. For Cluster, c belongs L
4. For nodes belonging to Cluster (c)
5. Define the conditional values for each PCH and new PCH
6. Define the Conditional values for SCH also
7. Assign the node condition to SCH and apply it to the broadcast of new messages
8. Check for empty messages. If true, iterate all 1-7 else close.

Algorithm 2: Intracluster routing algorithm

Input1: Initiate the number of nodes as n

Output: Create a cluster of Nodes on X and Y directions which are PCH, SCH and other affected nodes

1. Design messages for each set of Structures governing the Network as S and its distances from the sink.
2. Iterate the feature from all criteria n belongs to L1 for each node
3. Define minimum and maximum distances for the nodes to be operated
4. Assign a node for PCH
5. Define the clusters for the same number of nodes in S
6. Iterate for each set of clusters with SCH
7. Finally, estimate the conditional values for Node distances and their cluster distances.
8. Iterate for all S belongs to L
9. Broadcast the data with a message chosen with Cluster selected.

With features established based on the algorithm 1-2 this design effective implements congestion problem with routing methodology based on the structures (Fig-3 & Fig-4). Now to improvise this design for other network problems such as Cluster-head energy, minimum energy, this design should be optimized with other routing scheme's apart from congestion scenarios where such concepts are mentioned in section -II.

III. Optimizing Algorithms

A. Concept

The purpose of the design might be as simple as lowering manufacturing costs or increasing production efficiency. An optimization algorithm is a method carried out repeatedly by comparing several solutions until an optimal or satisfying answer is discovered. Optimization became an element of computer-aided design operations with the introduction of computers. Today's optimization algorithms are divided into two categories:

1) Deterministic
2) Stochastic

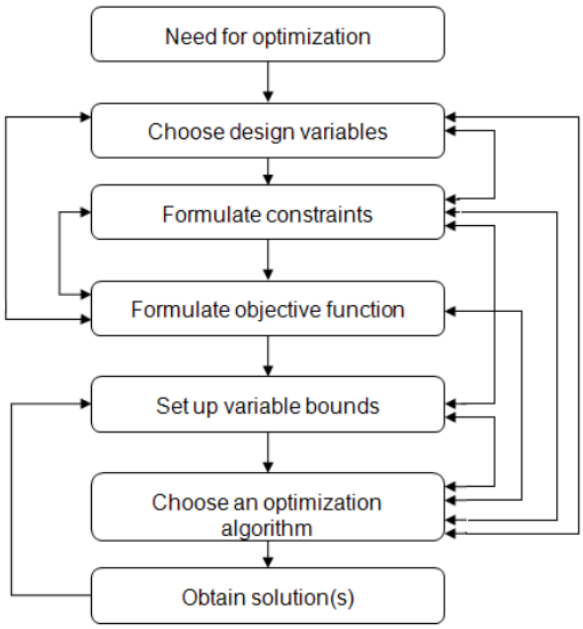


Fig. 5. Representing the Optimization Flow Diagram.

Our design improvises with the stochastic model on each Network parametric considered for optimizing. Since the optimization with energy-efficient is the prime scenario of the Node energy estimated and the node cluster energy modelled. In Fig 5 most specific feature is estimated with different variables and objective function with an optimization solution with solutions as fitness function values.

B. Link State Optimization

1) Initialize the design with a flag and graph where each node is utilized with maximum value.

2) Estimation on each tree graph is generated by selecting different MPR's based on Dijkstra's distance. We initiate the different scenario on the model where optimization is observed with minimum distance between each node estimated for selection of MPR's.

3) Optimizing the threshold values for each selected MPR's is generated.

4) Hopping on iterations changes would suffice the selection of MPR's selection onto the minimum energy model for given criteria.

Since, in section -I only congestion problem is illustrated with different routing structures on area implementation have been analyzed. While in section -II we improvise a conditional stochastic model with inequality criteria where this algorithm have been improvised to provide an effective solution for improved routing model with energy minimization problem as mention section -III and section -IV.

IV. Proposed Model

A. Concept

Network parametric criteria with specific features of design modules as proposed with bitrate, Cluster-head energy, minimum energy and bandwidth usage. In this design, three features of the Network parametric have been introduced with the distance formulation as proposed with the hybrid routing protocol. The figure shows the optimum solution for all the parametric features with area modelling hybrid routing and, finally, energy minimization. These features are estimated with proposed algorithms SCI and HPTTA for energy inequality estimation for minimum energy solutions for each iteration.

This design features the stochastic model on the Network parametric to improve the performance factors for each iteration set. Fig 6 describes the importance of energy optimization with node and Routing have become the crucial aspect of the design for estimating different network parametric criteria. We propose a pervasive stochastic model to ensure the node and route optimal feature for every case of observed energy values. The Routing would improvise a distance feature model for each set of selected nodes from the OLS algorithm, and an improved distance approach with SCI inequality is applied for improved energy minimization.

B. Area Set up for Topology Control

In figure 7a, we improvise a novel design scenario that effectively calculates the distances with R1, R2, R3 and R4 from the centre as d1, d2, d3 and d4. The adequate distances are formulated in section-3 for each region (1-4). The region in figure 7a with a star in-between represents the optimal regional transform for each tangential point as mentioned below:

Let p_i $(i: 1$ *to* $4)$*be the tangential points for star region considered*, where arc R1-R2 and arc R1-R4 will implicate the point P1 similarly for other arcs p2, p3, and p4 are represented.

Let E_j*being the edges of the star connected to circles with points* p_i $i = 1:4$, here j varies from 1 to 10. Even values of edge points are the critical points for which the tangential transform for the distances are applied.

$$\sum_{i=1}^{N} \sqrt{(D_{xi}^2 + D_{yi}^2) * \cos(\emptyset_i)} - \sqrt{(D_{xi}^2 + D_{yi}^2) * \sin(\emptyset_i)} \quad (3)$$

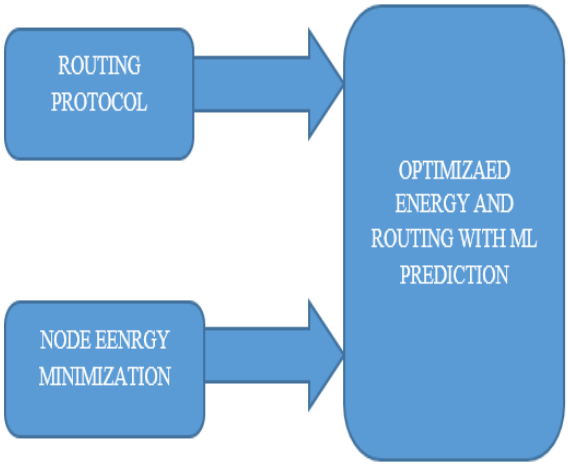
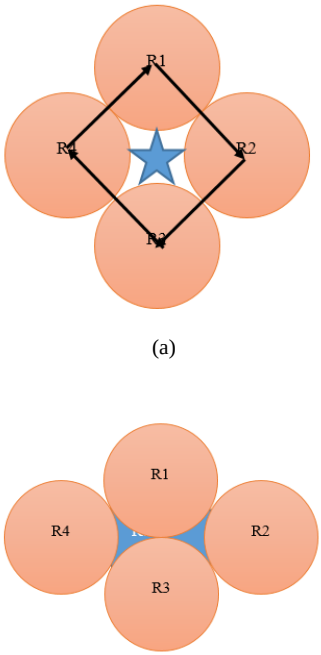


Fig. 6. Representing Optimization of Routing and Energy Minimization Block Diagram.



(a)

(b)

Fig. 7. (a) Representing the Pervasive Tangential Transform Model (b) Region Tangential with Projection Transform.

C. Hybrid Routing Model

With the Routing of hybrid feature, our design implicates the OLS algorithm as a proactive model and hybrid Manhattan distance feature for each OLS model, estimating the node minimum values for each node considered. The OLS algorithm improvises the design of MPR's multipoint relays, which are estimated with minimum values of distance based on the Hybrid Distance Formulations:

$$Dx1 = \sqrt{x_i \cos(45)} + \sqrt{y_i \sin(45)} \quad (4)$$

$$Dx2 = \sqrt{x_i \cos(15)} - \sqrt{y_i \sin(75)} \quad (5)$$

$$Dx = \min(Dx1, Dx2) \quad (6)$$

$$Final_{dist} = \sqrt{Dx1^2 + Dx2^2} \quad (7)$$

$$dist_{xy} = \frac{x_i}{2*pi*final_{dist}} - y_i\sqrt{2 * pi * Final_{dist}} \quad (8)$$

V. Energy-Efficient Formulations

A. Node Energy Formulations

We improvise an optimal solution for implicating Node estimations for hybrid routing as proposed, using formulations as mentioned for each set of nodes and selected clusters as:

This design for energy equation is measured with hybrid routing protocol as mentioned for topology Model:

$$S(i) = \sum_{j=1}^{K} \sum_{i=1}^{N} \mu * Fx(i,j) + \varphi * Fz(i,j) \quad (9)$$

The μ, φ The functionality of the S represents the design solution of which nodes appear at a given timing aspect where each set of the design parametric are considered with active and dead cells from the equation.

$$Fx(i > j) = \sum_{i=1}^{N} (n_i * Dmin(i) + \sigma * S(i)) \quad (10)$$

$$Fz(i \leq j) = \sum_{i=1}^{N} (n_i * D_{avg}(i)) \quad (11)$$

F represents the solution model for where all the active and alive nodes in cell regions are established with equation 1. n_i represents the number of active nodes and σ being the best predicted based on ATGF for all iterations mentioned in the above algorithms.

$$PE_{SCI} = F(i > k) + F(i \leq j) \quad (12)$$

$$PE_{Head\_cluster} = \gamma * W_i * D_{min} + \mu SCI(i) + E(i) \quad (13)$$

Here $E(i)$ represents the entropy of each selected feature on node localization selected.

Hence the total Network energy estimated is:

$$PE_T = PE_{Head\_cluster} + PE_{SCI} \quad (14)$$

Here $\gamma, \mu, \sigma,$ estimated probabilities for the optimized values for the best solution are modified with the SCI algorithm probabilities for each random variable estimation. The overall estimated values implicate the least energy simulated parametric for iteration chosen.

B. SCI Algorithm for Optimization of Minimum Energy

Let there be two variables, X and Y, the random probability for each node distance and energy optimization, on selecting optimal distance search and link-state protocol as mentioned in section 4.

$$P\left(\frac{X}{D_{mini}}\right) = P(X)\, P(X \cap D_{mini})/P(D_{mini}) \quad (15)$$

$$P\left(\frac{Y}{E_{mini}}\right) = P(Y)\, P(Y \cap E_{mini})/P(E_{mini}) \quad (16)$$

Since the conditionality of the stochastic model, $P(E_{mini}/D_{mini})$ is a conditional value for each Random variable X and Y for distance and energy minimization.

$$P\left(\frac{E_{mini}}{D_{mini}}\right) = P(E_{mini}) * \frac{P(E_{mini} \cap D_{mini})}{P(D_{mini})} \quad (17)$$

For $X \in D_{mini}$ and $Y \in E_{mini}$ hence,

$$P(E \cap D) = P(E) + P(X \cap D) + P\left(\frac{X}{D_{mini}}\right) * \frac{P(D)}{P(X)} + P(D_{mini}) + P\left(\frac{Y}{E_{mini}}\right) * \frac{P(E)}{P(Y)} + P(Y \cap E) \quad (18)$$

$$\text{Hence } P(E)_{min} \approx P\left(\frac{X}{D_{mini}}\right) * \frac{P(D)}{P(X)} + P(D_{mini}) + P\left(\frac{Y}{E_{mini}}\right) * \frac{P(E)}{P(Y)} \quad (19)$$

VI. Results and Discussion

A. Output Simulations

In this design, we have implicated the different features of the Network with different nodes and its multipoint routes estimated from the hybrid routing area transform model. The simulations are estimated with these nodes, and an optimal solution for the energy minimization and 1.8-1.9 bit rate is observed for the two algorithms proposed as SCI and HTTA from equation 3.

In figure 8, energy values are plotted with the minimum number of nodes in one direction. The minimum values for each set of iterations are observed with -220dB for 50X50 net, -238dB for 100X100 and -220dB for 150X150. Formulations from 1-19 are implemented for each setup iteration and compared with minimum energy values, bit rate and finally, the error values for each random factor chosen for optimizing the data as mentioned in figure 9-11.

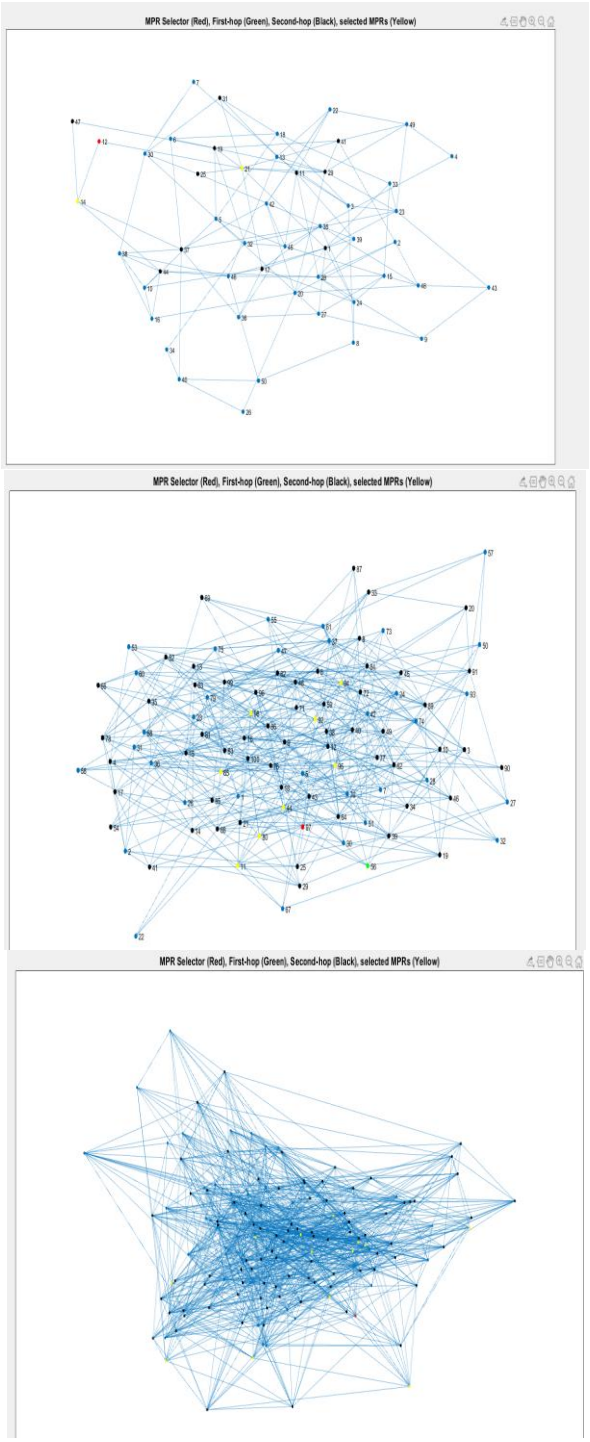


Fig. 8. Representing the WSN Structure for (a) 50X50 Nodes, (b) 100X100 Nodes (c) 150x150 Nodes.

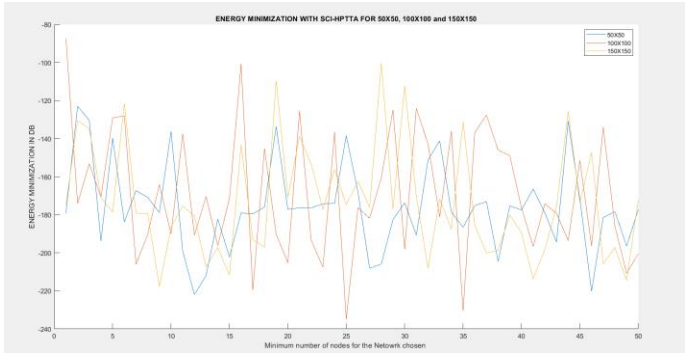


Fig. 9. Representing the Energy Values for (a) 50X50, (b) 100X100, (c) 150X150.

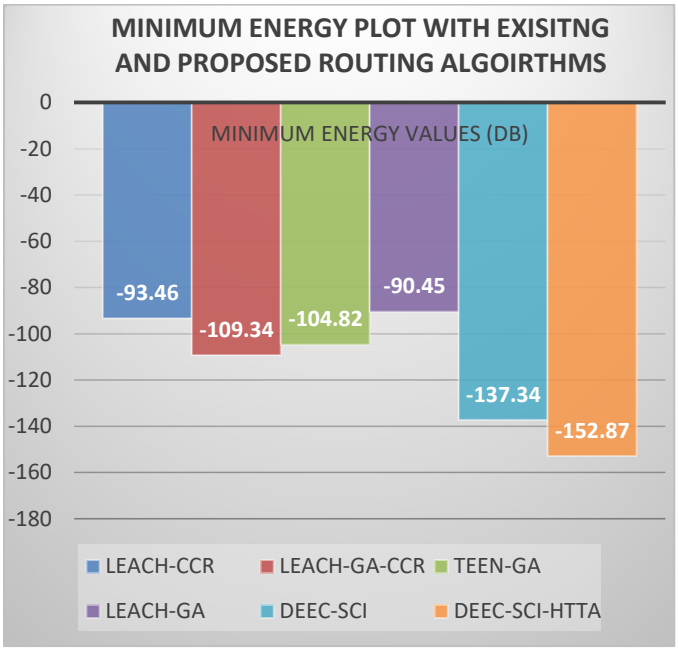


Fig. 10. Representing the Minimum Energy Values for Existing and Proposed Algorithms.

The minimum energy values of each algorithm mention with LEACH {CCR, GA, GA-CCR}, DEEC {SCI, SCI-HTTA proposed algorithm}, TEEN-GA are estimated with references [21,22]. The DEEC-SCI and DEEC-SCI-HTTA algorithm have been improvised on different solutions with each estimate on the equations from (15) and (18).

In Fig-11 and Fig-12, our design depicts with overall bit rate estimation for each transmitted data from the sender to the destination. To increase the bit of the transmission our design utilizes hybrid routing protocol with as proposed with area estimation structure. These area estimation analyses the correct distance from each node and its transmitter data to the send via SCI algorithm. Finally, the error optimization with hybrid routing scheme is analysed and implicated on the design feature. The error values are implicated with optimization algorithm as Stochastic model with SCI algorithm.

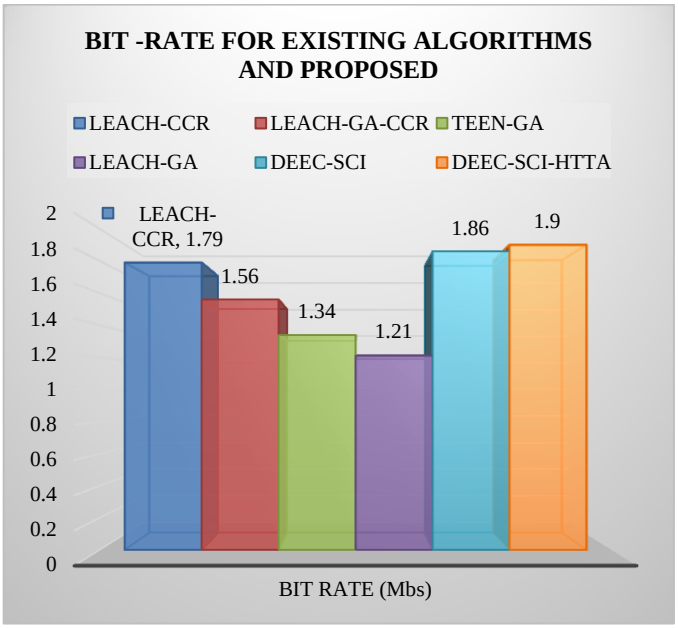


Fig. 11. Representing the Bit Rate Values for Existing and Proposed Algorithms.

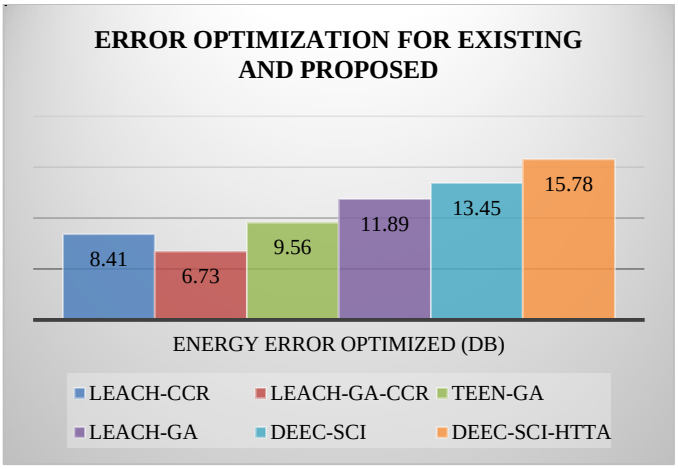


Fig. 12. Representing the Error Values in dB for Existing and Proposed Algorithms.

B. Tabulations

The Table 1a) implicates on different LEACH protocols from Existing model CCR referring to Congestion control route. The Fig-3 and Fig-4are the prime features for provision of routing with and without congestion for set of nodes as energy minimization. GA is other optimization algorithm considered for route minimization and its distance optimizing on each iteration for every node considered. Hence, Algorithm-2 from Table 1a) provisions best outcome for energy, bitrate and error in DB while others are less. So, in table-1b) is implemented based on the proposed algorithms as mention in section-III and IV encapsulating the different propositions of the energy models and its analysis for distance routes and node clusters minimizations. This results overall improved feature more than 12% on each of the parameter considered.

TABLE I. REPRESENTS THE DIFFERENT ALGORITHMS IMPLICATED WITH LEACH AND TEEN

SNO	Comparison table			
	Existing Algorithms	***Min-Energy***	***Bitrate***	***Error in dB***
1	LEACH-CCR	-93.46	1.79	8.41
2	LEACH-GA-CCR	-109.34	1.56	6.73
3	TEEN-GA	-104.82	1.34	9.56
4	LEACH-GA	-90.45	1.21	11.89

(B) REPRESENTS THE PROPOSED ALGORITHMS DEEC-SCI AND SCI-HTTA

SNO	Comparison table			
	Proposed Algorithms	***Min-Energy***	***Bitrate***	***Error in dB***
1	DEEC-SCI	-137.34	1.86	13.45
2	DEEC-SCI-HTTA	-152.87	1.9	15.78

VII. CONCLUSION

In WSN applications, power consumption is a common difficulty regardless of the task. To maintain the Network's overall efficiency, this design model with SCI and HTTA is implemented with energy minimization is required. This work intended with Routing and its area-based design model with tangential transform for different Network parametric as SCI-HTTP model with DEEC is proposed. The optimized features on performance parametric suggest an outstanding improvement on minimum energy values for the proposed two algorithms. In figure 2, we have implicated the DEEC-SCI and DEEC -SCI-HTTA features for the proposed model. These few algorithms are being considered with the proposed improvement feature on Routing speed characteristics. From section -6, all tabulations and plotting features are represented via Matlab-2019b. The practical value minimum value for each DEEC-SCI-HTTA is observed in terms of -200 to -300 dB for a single iteration.

SCOPE:

1) To estimate heterogenous network for optimizing coverage problem

2) Improvise a Novel scheme on optimal energy nodes and its provision on hardware model for WSN network.

REFERENCES

[1] K Abdul Basith, T.N. Shankar, “Hybrid state analysis with improved firefly optimized linear congestion models of WSNs for DDOS & CRA attacks”, URL link: Hybrid state analysis with improved firefly optimized linear congestion models of WSNs for DDOS & CRA attacks [PeerJ], pp-11-14,DOP:2022-01-27.

[2] Muhammad Atif Ur Rehman; Rehmat Ullah; Byung-Seo Kim; Boubakr Nour; Spyridon Mastorakis, “ CCIC-WSN: An Architecture for Single-Channel Cluster-Based Information-Centric Wireless Sensor Networks”, IEEE Internet of Things Journal (Volume: 8, Issue: 9, May 1, 1 2021) Page(s): 7661 – 7675 Electronic ISSN: 2327-4662.

[3] F. Fernando Jurado-Lasso; Ken Clarke; Ampalavanapillai Nirmalathas, “Performance Analysis of Software-Defined Multihop Wireless Sensor Networks”, IEEE Systems Journal (Volume: 14, Issue: 4, Dec. 2020).

[4] T. M. Behera, S. K. Mohapatra, U. C. Samal, M. S. Khan, M. Daneshmand and A. H. Gandomi, "Residual energy-based cluster-head selection in WSNs for IoT application", IEEE Internet Things J., vol. 6, no. 3, pp. 5132-5139, Jun. 2019.

[5] M. Ndiaye, G. P. Hancke and A. M. Abu-Mahfouz, "Software-defined networking for improved wireless sensor network management: A survey", Sensors, vol. 17, no. 5:1031, pp. 1-32, 2017.

[6] O. Salem, A. Serhrouchni, A. Mehaoua and R. Boutaba, "Event detection in wireless body area networks using Kalman filter and power divergence", IEEE Trans. Netw. Service Manag., vol. 15, no. 3, pp. 1018-1034, Sep. 2018.

[7] C. Chen, J. Yan, N. Lu, Y. Wang, X. Yang and X. Guan, "Ubiquitous monitoring for industrial cyber-physical systems over relay-assisted wireless sensor networks", IEEE Trans. Emerg. Topics Comput., vol. 3, no. 3, pp. 352-362, Sep. 2015.

[8] S. Misra, S. Bera, M. Achuthananda, S. K. Pal and M. S. Obaidat, "Situation-aware protocol switching in software-defined wireless sensor network systems", IEEE Syst. J., vol. 12, no. 3, pp. 2353-2360, Sep. 2018.

[9] M. Ndiaye, G. P. Hancke and A. M. Abu-Mahfouz, "Software-defined networking for improved wireless sensor network management: A survey", Sensors, vol. 17, no. 5:1031, pp. 1-32, 2017.

[10] K. Latif, N. Javaid, M. N. Saqib, Z. A. Khan and N. Alrajeh, "Energy consumption model for density controlled divide-and-rule scheme for energy-efficient routing in wireless sensor networks", Int. J. Ad Hoc Ubiquitous Comput., vol. 21, no. 2, pp. 130-139, 2016.

[11] G. Ateniese et al., "Low-cost standard signatures for energy-harvesting wireless sensor networks", ACM Trans. Embedded Comput. Syst., vol. 16, no. 3, 2017.

[12] F. Wu, L. Xu, S. Kumari and X. Li, "A privacy-preserving and provable user authentication scheme for wireless sensor networks based on internet of things security", J. Ambient Intell. Humanized Comput., vol. 8, no. 1, pp. 101-116, 2017.

[13] H. Zhang and Z. Pan, "Cross-voting SVM method for multiple vehicle classification in wireless sensor networks", Sensors, vol. 18, no. 9, pp. 3108, 2018.

[14] N. Jain, S. Verma and M. Kumar, "Adaptive locally linear embedding for node localization in sensor networks", IEEE Sensors J., vol. 17, no. 9, pp. 2949-2956, May 2017.

[15] A. Singh and S. Verma, "Graph Laplacian regularization with Procrustes analysis for sensor node localization", IEEE Sensors J., vol. 17, no. 16, pp. 5367-5376, Aug. 2017.

[16] W. Heinzelman, A. P. Chandrakasan, H. Balakrishnan, and A. C. Smith, "Application-specific protocol architectures for wireless networks," PhD dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2000.

[17] V. Loscri, G. Morabito, and S. Marano, "A two-levels hierarchy for low-energy adaptive clustering hierarchy (TL-LEACH)," in Proc. IEEE Veh. Technol. Conf., Sep. 2005, vol. 62, no. 3, pp. 1809–1813.

[18] L. Lijun, W. Hongtao, and C. Peng, "Discuss in round rotation policy of hierarchical route in wireless sensor networks," Proc. Int. Conf. Wireless Commun., Netw. Mobile Comput., Sep. 2006, pp. 1–5.

[19] F. Xiangning and S. Yulin, ''Improvement on LEACH protocol of wireless sensor network,'' in Proc. Int. Conf. Sensor Technol. Appl., Oct. 2007, pp. 260–264.

[20] H. Junping, J. Yuhui, and D. Liang, ''A time-based cluster-head selection algorithm for LEACH,'' in Proc. IEEE Symp. Comput. Commun., Jul. 2008, pp. 1172–1176.

[21] M. S. Ali, T. Dey, and R. Biswas, ''ALEACH: Advanced LEACH routing protocol for wireless microsensor networks,'' in Proc. Int. Conf. Elect. Comput. Eng., Dec. 2008, pp. 909–914.

[22] W. Wang, Q. Wang, W. Luo, M. Sheng, W. Wu, and L. Hao, ''LEACHH: An improved routing protocol for collaborative sensing networks,'' in Proc. Int. Conf. Wireless Commun. Signal Process., Nov. 2009, pp. 1–5.

[23] G. Yi, S. Guiling, L. Weixiang, and P. Yong, ''Recluster-LEACH: A recluster control algorithm based on density for wireless sensor network,'' in Proc. 2nd Int. Conf. Power Electron. Intell. Transp. Syst. (PEITS), vol. 3, Dec. 2009, pp. 198–202.

[24] A. Yektaparast, F.-H. Nabavi, and A. Sarmast, ''An improvement on LEACH protocol (Cell-LEACH),'' in Proc. 14th Int. Conf. Adv. Commun. Technol. (ICACT), Feb. 2012, pp. 992–996.

[25] G. N. Basavaraj and C. D. Jaidhar, ''H-LEACH protocol with modified cluster head selection for WSN,'' in Proc. Int. Conf. Smart Technol. Smart Nation, Aug. 2017, pp. 30–33.

[26] A. Razaque, M. Abdulgader, C. Joshi, F. Amsaad, and M. Chauhan, "P-LEACH: Energy-efficient routing protocol for wireless sensor networks," in Proc. IEEE Long Island Syst., Appl. Technol. Conf. (LISAT), Apr. 2016, pp. 1–5.

[27] Uma Maheswari Durairaj; Sudha Selvaraj, "Two-Level Clustering and Routing Algorithms to Prolong the Lifetime of Wind Farm-Based WSN", IEEE Sensors Journal (Volume: 21, Issue: 1, Jan.1, 1 2021), DOI: 10.1109/JSEN.2020.3015734.

[28] Tarunpreet Kaur; Dilip Kumar, "MACO-QCR: Multi-Objective ACO-Based QoS-Aware Cross-Layer Routing Protocols in WSN", IEEE Sensors Journal (Volume: 21, Issue: 5, March1, 1 2021), Electronic ISSN: 1558-1748.

[29] Xiuwen Fu; Yongsheng Yang; Octavian Postolache, "Sustainable Multipath Routing Protocol for Multi-Sink Wireless Sensor Networks in Harsh Environments", IEEE Transactions on Sustainable Computing (Volume: 6, Issue: 1, Jan.-March 1 2021), Electronic ISSN: 2377-3782.

[30] Shaha Al-Otaibi, Amal Al-Rasheed,Romany F. Mansour, Gyanendra Prasad Joshi, Woong Cho, " Hybridization of Metaheuristic Algorithm for Dynamic Cluster-Based Routing Protocol in Wireless Sensor Networksx", IEEE Access (Volume: 9), Date of Publication: 08 June 2021.

FNU-BiCNN: Fake News and Fake URL Detection using Bi-CNN

R.Sandrilla[1]
Research Scholar, Department of Computer Science
Periyar University, Salem
Tamilnadu, India

M.Savitha Devi[2]
Assistant Professor, Department of Computer Science
Government Arts and Science College, Harur
Dharmapuri DT, Tamilnadu, India

***Abstract*—Fake news (FN) has become a big problem in today's world, recognition partly to the widespread use of social media. A wide variety of news organizations and news websites post their stories on social media. It is important to verify that the information posted is genuine and obtained from reputable sources. The intensity and sincerity of internet news cannot be quantified completely and remains a challenge. We present an FNU-BiCNN model for identifying FN and fake URLs in this study by analyzing the correctness of a report and predicting its validity. Stop words and stem words with NLTK characteristics were employed during data pre-processing. Following that, we compute the TF-IDF using LSTM, batch normalization, and dense. The WORDNET Lemmatizer is used to choose the features. Bi-LSTM with ARIMA and CNN are used to train the datasets, and various machine learning techniques are used to classify them. By deriving credibility ratings from textual data, this model develops an ensemble strategy for concurrently learning the depictions of news stories, authors, and titles. To achieve greater accuracy while using Voting ensemble classifier and compared with several machine learning algorithms such as SVM, DT, RF, KNN, and Naive Bayes were tried, and it was discovered that the voting ensemble classifier achieved the highest accuracy of 99.99%. Classifiers' accuracy, recall, and F1-Score were used to assess their performance and efficacy.**

***Keywords*—Bi-LSTM; CNN; WORDNET; machine learning; fake news and URL; ARIMA**

LIST OF ABBREVIATIONS

FN-Fake news

FNU-BiCNN- Fake News and Fake URL Detection Using Bi-CNN

SVM- Support Vector Machine

DT-Decision Tree

RF-Random Forest

KNN- K-Nearest Neighbors

NB- Naive Bayes

CNN- Convolutional Neural Network

LSTM- long short-term memory network

ARIMA- Autoregressive Integrated Moving Average

URL-Uniform Resource Locators

MSSE- Minimum Sum of Squared Errors

BP- Back Propagation

TF- Term Frequency

IDF- Inverse Document Frequency

POS-Part of Speech

RNN- recurrent neural network

NLTK- Natural Language Toolkit

I. INTRODUCTION

Recent years have seen a rapid rise in the popularity of social networking sites due to greater media coverage. Rather than traditional media, social networking platforms are the preferred news source for many people [1]. Users of social networks may engage with individuals who share their interests and ideas. It's questionable, though, about the quality of the news. This media outlet disseminates false material in the style of news stories [2]. People and society might be adversely affected by its widespread use. FN is information that has been created solely to mislead the public. It is impossible to accurately measure the reliability of information posted on social media networks [3]. To overcome the problem above, a standardized solution is needed [4].

Numerous dangerous consequences might result from our culture's exponential growth of FN. First, FN alters how people view and respond to legitimate news. Second, the proliferation of FN [11] would consumers' faith in the media, make them distrustful, and jeopardize the news medium's legitimacy [7] [8]. Third, deliberate FN persuades people to accept skewed and manufactured tales [10].

There is lot of reasons to choosing the FN detection: First, true or untrue stories greatly influence a country's elections, such as in India, where 45% of voters believe fact-checking groups exist. [11] [12]. Most WhatsApp users in India prefer to accept transmitted information without checking it, which is the second-largest population in the world [13-17]. On the one hand, social media businesses are faced with the enormity of their enterprise as they consider the possible exploitation of its base. Examples of FN are photoshopped images, client-created content, or caricature accounts. Second, there is mounting evidence that customers have behaved bizarrely in response to news that was subsequently shown to be false. One recent example is the propagation of the new corona virus, which was propagated by false claims about the virus's origin, biology, and behavior. The situation deteriorated as more individuals became aware of the fabricated information online. Finding such news online is a difficult Endeavour.

Other types of FN include stories intended to appeal to a specific group or association and stories that offer a scientific or affordable explanation for an unresolved issue, leading to the spread of incorrect information. FN detection and Fake URL detection bring new and tough difficulties due to the aspects above [18-21].

For detecting malicious URL has merging a trust computational model with a collection of URL-based characteristics. And Malicious URL detection has used Bayesian learning and Dempster–Shafer theory to assess the credibility of tweets and it has only 95% accuracy rate [16].

Internet news items may automatically detect FN and URL information in internet news items using two new datasets. Data sets that have been pre-processed are utilized to distinguish between fake and legitimate news. ARIMA, Bi-LSTM, and Convolutional Neural Networks (CNN) deep learning algorithms were utilized for training the datasets. Our last step is using ensemble learning, in which we combine many classifiers to improve a model's ability to classify and approximate new data accurately. Base classifiers include Support Vector Machine (SVM), KNN, Naive Bayes [NB], Decision Tree (DT), and Random Forest (RF). Together, these two classification models provide a better estimate and a classifier that beats all others in terms of accuracy and predictability. By combining numerous models and then averaging them to generate a final model, this method helps limit the danger of a poor-performing classifier.

The following are the primary goals and accomplishments:

A strategy for spotting FN across several sources:

a) Extract top-level text features from actual and FN articles using TF-IDF and WORDNET.

b) URL characteristics may be extracted by looking at the domain name (domain)

c) The on-site URL feature may be used to estimate the multi-source trustworthiness score by combining text-based characteristics and multi-source credibility ratings to estimate news credibility.

A. Our Contributions

Multiple instances of text classification using both supervised and unsupervised learning methods have been seen in the present FN corpus. However, the majority of the research focuses on certain datasets or topics, most notably the realm of politics. As a consequence, the algorithm trained on a specific kind of content performs optimally when exposed to articles from various areas. Our research examines many textual features that may be utilized to detect false from genuine information. We use these qualities to train a mix of distinct machine learning algorithms utilizing voting ensemble approaches that have not been extensively investigated in the present literature. Voting ensemble learners have been shown to be beneficial in a broad number of applications due to their propensity to lower error rates. These strategies assist the effective and efficient training of various machine learning algorithms. Additionally, we performed extensive experiments on two publicly accessible real-world datasets as Fake news and URL datasets.

The rest of the article is arranged in the following manner: Section 2 deals with existing research methods for FN and URL detection. Section 3 explains how the FNU-BiCNN architecture works and how it is put into action. Section 4 presents the results of the research into the FNU-BiCNN framework. Conclusions and future research are discussed in Section 5.

II. Background Study

Agarwal, A., & Dixit, A. in [1] instead of studying a single approach, the author employed ensemble learning. The average accuracy score discovered was 85%, which is 15% better than the accuracy of the worst-performing KNN model. Additionally, the authors utilized just a percent of the information even with the supplied dataset and values. The remainder of the data was inadequate and did not give further distinguishing characteristics between fake and authentic news.

In Ahmed, A. A., & Abdullah, N. A. [2], the URLs of online pages may be used to identify phishing websites. The proposed method might distinguish between legitimate and counterfeit websites by looking at the Uniform Resource Locators (URLs) of suspicious web pages (URLs). URLs are examined for a variety of characteristics to identify phishing sites. The identified attacks are reported to the proper authorities to prevent such incidents.

In Birunda, S. S., & Devi, R. K. [3], a novel score-based framework for detecting FN from multiple sources has been developed. Using the TF-IDF approach, the top actual and false characteristics were retrieved from news articles. The Credibility Score of the sources was determined using the Site URL attributes provided from the source. To determine the news's dependability, the retrieved text-based characteristics and the multi-source Credibility Score were combined. The suggested framework's efficacy and practicality are assessed and compared to different classifiers.

Cheng, W. et al. in [5] for integrated forecasting models, presents a novel weighting approach for MSSE (Minimum Sum of Squared Errors) models that combine ARIMA (Autoregressive Integrated Moving Average) time-series models with BP (Back Propagation) neural networks.

Granik, M., & Mesyura, V. in [6] shows how to use a naive Bayes classifier to identify FN quickly. This strategy was developed and tested using a computerized system using Facebook data collection. For a quite simple model, the authors achieved a classification accuracy of around 74% on their test set.

In Jayasiriwardene, T. D., & Ganegoda, G. U. [9] to gather data to detect FN stories, this article demonstrates a method for extracting keywords from a particular tweet's body text. The proposed approach uses Stanford core NLP, POS tagging, and TF-IDF statistical techniques to identify keywords. The Wordnet lexical database was used to find synonyms, and Ginsim and word2vector may be used in combination to assess how related two words are. Using a bi-gram technique, keywords are created to increase news retrieval accuracy and efficiency. A list of the most relevant news tweets regarding the claimed tweet is compiled using the extracted keywords.

In Mansouri, R. et al. [13], use of semi-supervised linear discriminant analysis and CNNs are described in the following article to detect FN. With the addition of unlabeled datasets, it was necessary to increase training data. The proposed method alters LDA to provide a semi-supervised estimation of classes.

Qazi, M. et al. [15] discusses a novel approach for detecting fraudulent or legitimate news. For detecting purposes, liar data is employed. The literature review reveals that several machine learning-based detection approaches identify FN. However, these models lack performance. The authors attempt to increase performance by developing a transformer model based on the attention process.

In Rout, R. R., et al. [16], authors present a LA-MSBD method for detecting malicious social bots by merging a trust computational model with a collection of URL-based characteristics (MSBD). Additionally, the authors are used Bayesian learning and Dempster–Shafer theory to assess the credibility of tweets. MSBD has a 95% accuracy rate.

In Seo, Y., & Jeong, C.S. [19], the authors validated the proposed model using two distinct dataset types. To begin, when the sequence-to-sequence learning model is utilized, the parallel corpus dataset is used as an input for creating the sentences. Second, to modify CNN news articles provided by DeepMind and construct different sorts of propositions in order to test inference performance.

In Vogel, I., & Meghana, M. [20] as a first step in limiting the transmission of false news among online users, the authors presented three distinct ways for automatically detecting suspected FN spreaders on social media. The authors utilized the PAN 2020 author profile corpus and performed a variety of multilingual learning tests. To assess a variety of handmade and machine-learned qualities, the majority of which are language-independent? The characteristics were retrieved and their significance in the detection job was determined.

In Zhi, X. et al. [21], the authors present a unique model based on CNN-LSTM that incorporates news organizations, comments, sources, and market data in this study. The findings demonstrate that our strategy beats previous work that manually extracts features or criteria.

III. Proposed Framework

This section adds to the comprehensive approach of the proposed 'FNU-BiCNN' framework. The news data are first gathered, filtered, and pre-processed. Then, text-based elements are retrieved to determine the reliability of news items and URL information. Additionally, an ensemble technique is presented to use the source's credibility score and text-based elements to determine the authenticity of a news story and URLs. The FNU-BiCNN framework's pipeline is shown in Fig. 1.

A. Problem Statement

If a news article's content and its credibility score are both proved to be false (false) and genuine (true), otherwise, it is considered to be a fake.

Allowing for the fact that an infinite number of news stories originate from numerous sources, we designate this dataset as $D = D_1, D_2..., D_{ns}$. Y = "Real, Fake" is the dataset's classification code. The new unlabeled news article's degree of fakeness may be predicted using the dataset D, news articles from numerous sources (s), and the class labels Y.

B. Dataset

www.Kaggle.com's experimental dataset comprises 23,000 news stories that are either false or authentic. The data collection is 100 MB in size. This dataset contains information on the source of the news stories, the date of publication, the author, the title, and the text. True and fake datasets are used for training and testing purposes. 976 false and 598 actual news pieces are utilized for training purposes out of a total of 1574 articles. 241 fake samples and 152 authentic samples were found in 393 news stories. The difficulty of detecting FN articles is one of ML classification. There are 1217 FN pieces and 750 legitimate news articles shown in Fig. 2. More than 1500 URL links with descriptions were gathered from kaggle.com for the URL dataset. The dataset is 50 MB in size. The link source, title, and description are all included in this dataset. As a data source for training and testing, this dataset is divided into real URLs and fake URLs.

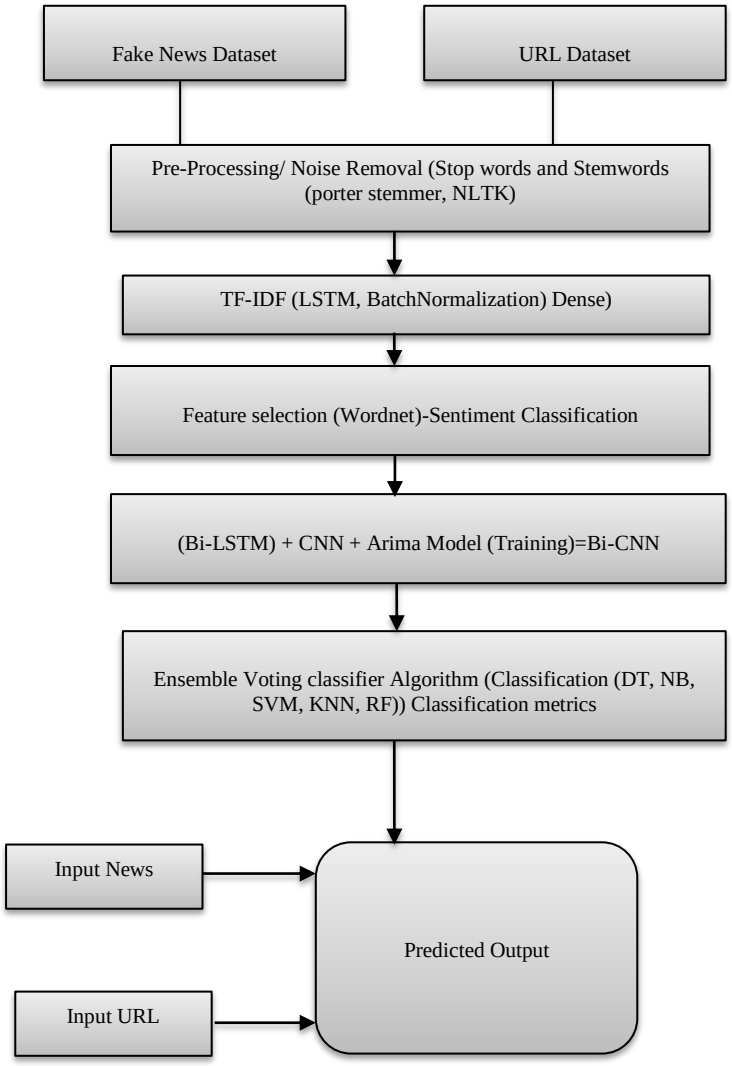


Fig. 1. FNU-BiCNN Model for Detection Fake News and URL.

C. Preprocessing

Processing raw data into a machine-readable format is known as pre-processing in the context of data mining. Many text pre-processing operations were carried on the news and URL datasets. While working toward these goals, the Keras library's algorithms were employed for character conversion to lowercase letters, stopword removal, stemming, and tokenization. Stopword, such as 'of,' 'the,' 'and,' 'an,' and similar words, are often found in texts used in this work. Eliminating stopword speeds up processing and frees up space formerly used by the useless words listed above. Words with similar meanings often occur in the text, such as games and sport. The strength of shortness is in the removal of unnecessary words. This is known as stemming, and it's done using the Porter stemmer technique from the NLTK open-source version.

The headlines' keyword density was reduced to 372 due to the aforementioned pre-processing steps. Keras' library's tokenizer function divided each headline into a vector of words. The text is first transformed into vectors using word embedding (word2vec). Finally, a vocabulary for the 5,000 unigram words present in headlines and article bodies is developed. The maximum headline length is used for all headlines. Padding is not applied to headlines that are shorter than the maximum length.

D. Feature Extraction

In our FNU-BiCNN approach, we analyze the extracted characteristic from the news and link it to the news's speaker or author. In addition, the linked author is given a credibility score depending on the number of fake and real stories he publishes on his website. A good credit score assures accuracy and trustworthiness. This is why our system approaches the problem in two ways: first, by extracting news text's semantics, and second, by giving a credibility score to each author. As a result, a vector matrix is formed that links the word list to each source. We turned to the sci-kit learn python libraries for feature extraction and selection. After selecting features using the bag of words and n-grams technique, we used TF-IDF weighting (Term Frequency-Inverse Document Frequency). Additionally, we used word2vec and POS tagging to extract features. There is a short discussion of each feature extraction model:

Stop words will be labeled, and the labels will be used to eliminate them. WordNet is used to generate semantic phrases related to the subject terms. WordNet is an API that can generate synonyms for a given term. The TF-IDF approach is used to calculate the weight of these words.

- Bag-of-words: It is a basic method for representing text data and extracting textual characteristics. This section tokenizes and counts the words associated with each observation.

$$Weight = TF * IDF \quad (1)$$

For a term *i* in document *j:*

$$W_{ij} = tf_{ij} X log(\frac{N}{df_i}) \quad (2)$$

tf_{ij}=number of occurrences of i in j

df_i=number of documents containing i

N=total number of documents

$$w^x + b = 0 \quad (3)$$

Where w is the weight vector and b are the bias

$$L(w) = \sum_1 \max(0, 1 - y_i[w^t x_i + |b|]) + \ |w||^2 \quad (4)$$

- The term "n-gram" refers to an uninterrupted sequence of n tokens or words. To get the most out of your n-grams, it's better to use a bag of words rather than just words.
- Information may be retrieved using the TF-IDF algorithm. This measure provides an accurate value when the token is commonly used in the document and frequently used in the corpus. Each word is given a word frequency score, with the more interesting ones receiving the highest marks.
- Word2vec: This method embeds semantically similar words adjacently in numeric vectors called embeddings.
- POS-Tagging: POS tagging is described as the act of associating a word with certain portions of speech depending on its context and meaning. It may be employed to resolve grammatical ambiguity or disambiguate word senses to establish the news's legitimacy.

Proposed Algorithm 1: Bi-CNN

Algorithm 1: Bi-CNN:
Input: News articles and URL datasets
Output: Fake or Real Data

Step 1: Pre-processing using NLTK[s] porter stemmer algorithm
Step 2: Extract Top features using TF-IDF using LSTM, Batch normalization, and dense
Step 3: Select Top features using TF-IDF and Bag of words
Step 4: If features are selected, then
Step 5: Using WORDNET sentiment analysis
Step 6: Else
Step 7: Repeat step 1
Step 8: Training data using ARIMA+BiLSTM+CNN
Step 9: Find the Training accuracy and loss
Step 10: Classification is done by ML Algorithms
Step 11: Find the Classification Metrics
Step 12: Create a pickle file
Step 13: Find the fake news and URLs
Step 14: End

E. Training Data

1) Arima model: The abbreviation ARIMA stands for Auto-Regressive Integrated Moving Average (ARIMA). When the lags of the stationarized series are included in the forecasting equation, it is referred to as an "autoregressive"

component. When the lags of prediction errors are used, "moving average" is used. When a time series must be differentiated to be deemed inactive, it is referred to as an "integrated" form of a stationary series. ARIMA models include random-walk and random-trend models, autoregressive models, and exponential smoothing models [12].

The notation "ARIMA (p, d, q)" represents an ARIMA model that is not seasonal, with p signifying the number of autoregressive components and d and q denoting the number of dependent variables.

The number of nonseasonal deviations necessary for stationarity is given by d in the prediction equation, whereas the number of delayed forecast errors is denoted by q.

The forecasting equation is constructed in the following manner. Let's start by referring to y as the d^{th} difference of Y, which equals eq (5,6,7):

$$if\ d = 0, y_t\ equals\ Y_t \tag{5}$$

$$if\ d = 1, y_t = Y_t - Y_t - 1 \tag{6}$$

$$if\ d\ is\ equal\ to\ 2, then\ Y_t = (Y_t - Y_t)2Y_t - 2 = Y_t - 2Y_t - 1 + Y_t - 2 = (Y_t - 1Y_t - 2) \tag{7}$$

a) FN Detection in Time Series: Because of the nature of the data, we cannot employ traditional FN detection algorithms to detect anomalies in time series data. The challenge of time series FN identification must be addressed separately from the other jobs.

The following are the actions that must be followed to detect FN in time-series data:

- Make a note of whether the FN data is moving or not. Make the FN data motionless by changing it too stationary if it isn't already.
- The study's findings fit a time series model to the pre-processed FN data.
- Calculate the observation's Squared Error for every observation in the data.
- Calculate your data's error threshold.
- We can label an observation as FN if the number of mistakes surpasses a specific threshold.

As previously established, time-series FN data is strictly sequential and prone to autocorrelation in distribution. Time-series models would be trained and utilized to discover the general behavior of the fictional data, and they would attempt to forecast the actual data using the fictitious data. If an observation is normal, the forecast will be as close to the true value as feasible; but, if an observation is an FN, the forecast will be as far from the true value as possible; hence, by studying the forecast errors, we may detect the FN in the data [5].

2) Bi-LSTM: A recurrent neural network (RNN) is the go-to choose when dealing with sequential input. It stores and uses just the most significant parts of the incoming data to predict the future output. Memory cells in the RNN keep track of the most important information from earlier inputs. Long-term dependence isn't taken into consideration either. The consequence is the creation of a particular RNN to deal with long-term dependency. It is known as the long short-term memory network (LSTM). Input (IG), output (OG), and forget gates are three of the gating principles used to do this (FG). (8)– (11) explain how the information flow (read, write, and reset) in the gradient is regulated by these gates in conjunction with the candidate hidden state (CHS), current state (CS), and hidden sequence (HS) (10). (12). Left-to-right or right-to-left are the only two input directions that the LSTM network accepts for processing. Consequently, gathering new information will be more difficult in the future. Since the original input sequence is being learned in both directions, Eq. employs a bidirectional LSTM in this case (13).

$$IG_t = \sigma(w_{IG}x_t + R_{IG}h_{t-1} + b_{IG}) \tag{8}$$

$$OG_t = \sigma(w_{OG}x_t + R_{OG}h_{t-1} + b_{OG}) \tag{9}$$

$$FG_t = \sigma(w_{FG}x_t + R_{FG}h_{t-1} + b_{FG}) \tag{10}$$

$$CHS_t = tanh(w_{CHS}x_t + R_{CHS}h_{t-1} + b_{CHS}) \tag{11}$$

$$CS_t = FG_t \text{ X } CS_{t-1} + IG_t \tag{12}$$

$$Y_t = V(H\ s_t : H\ S_t) \tag{13}$$

Where w_{IG}, w_{OG}, w_{FG}, and w_{CHS} are referring to the weight matrices of the current input x_t. R_{IG}, R_{OG}, R_{FG}, and R_{CHS} are referred to as the weight matrices of the previous state h_{t-1}. R_{IG}, R_{OG}, R_{FG}, and R_{CHS} are denoted as the bias value. Y_t represents the output of the forward LSTM and backward LSTM units.

3) Convolutional neural network: CNN: To summarize, CNN is a deep learning model that does very well in image categorization and automated natural language processing. When it comes to identifying higher-level traits, CNN has a distinctive structure. The primary processing unit of CNN is the convolutional layer, which uses matrix coefficients to identify features. There are several kernels or filters in this layer. These filters use an activation function called ReLU to help process a portion of the input sequence throughout the whole input data set (Rectified Linear Unit). As an extension of the corrected linear unit, the ReLU aims to remove negative values from the activation map by setting it to zero in most cases. It is, therefore, more effective than the sigmoid and the Tanh activation functions for solving the problem of the unseen gradient.

As a result of the convolutional approach, a fixed-sized word-level embedding may be obtained by first aggregating the local features generated by the neural network around each word in the neighboring word. As a result, CNN is used to model latent textual properties to identify FN. Let the j^{th} word in the news *i* denote as $x_{i,j} \in R^k$, which is a k-dimensional word embedding vector. Believe the maximum length of the news is n, s.t., the news has less than n words can be padded

as a series with length n. Hence, the overall news can be written as

$$X_{i,l:n}^{T1} = x_{i,1} \oplus x_{i,2} \oplus x_{i,3} \ldots \oplus x_{i,n} \quad (14)$$

This means that the news $X_{i,l:n}^{T1}$ is concatenated by every word. In this case, each news can be presented as a matrix. Then we use convolutional filters $\omega \in R^k$ to construct the new features. The feature c_i as follows:

$$c_i = f\left(w.X_{i,j:j+h-1}^{T1} + b\right) \quad (15)$$

Where, the b$\in R^k$. A max-pooling layer is applied to take the maximum feature map c. The maximum value is denoted as c=max{c}. Convolutional discoveries may be saved for FN detection in the max-pooling layer, increasing the model's durability by putting the pooling results into a fully connected layer.

4) Ensemble model for classification: SVM, DT, RF, KNN, and NB classification algorithms combine and use voting ensemble classifier algorithms. All of the code was written in Python. Python libraries are used in our source code: Keras, NLTK, NumPy, Pandas, Sklearn, and scikit. Algorithms were judged on their accuracy, precision, recall, and F-score, among other metrics.

a) Naïve Bayes is a machine learning method used to solve text categorization issues. Apart from that, it is quite simple to apply and extremely effective.

b) SVM (Support Vector Machine): When it comes to regression-based classification, an SVM is a common tool. Despite this, it is a common tool for resolving categorization issues. A point on an N-dimensional space, where N is the number of spaces, is often used to represent each piece of data. In addition, each element has its value, which indicates a certain location. Finally, we arrive at the hyperplane on which the data are grouped. Outliers may be excluded using the SVM algorithm while calculating the hyperplane.

c) KNN (K-Nearest Neighbors): It is one of the simplest machine learning models that adhere to the principles of supervised learning. It forecasts the similarities between the data for which the class is predicted and the data for which the class is already predicted. It classifies the new situation as a class, which is quite similar to a record or data. It is also applicable to regression and classification. It is, however, mostly utilized in categorization.

d) Random Forest Algorithm (RF): This is a well-known technique for supervised machine learning. Classification and regression are two possible applications. It uses ensemble learning, which combines the results of several classifiers to improve overall accuracy.

e) DT (Decision Tree): In classification, the DT algorithm is one of the most often used methods. C4.5 algorithm, which needs all data to be quantitative or categorical, is used. Continuous data will not be analyzed as a consequence of this decision. Pruning in DT may be accomplished in two ways. For instance, one method, termed "subtree replacement," proposes replacing nodes in a decision tree to minimize the number of tests in the convinced path. Typically, subtree raising has a negligible effect on decision tree models. Typically, there is no precise method to anticipate the utility of choice. However, it may be prudent to disable it if the induction method takes longer than expected due to the computational complexity of raising the subtree.

IV. Results and Discussion

CUDA is a technology developed by NVIDIA. Python offers a driver and runtime API for existing toolkits and libraries, simplifying GPU-accelerated computation for optimum performance while maintaining simplicity. The data set of FN material, which can be accessed here, and the news URLs, which can be viewed here, were pre-processed using the Porter stemmer algorithm and NLTK. The Natural Language Toolkit (NLTK) is a critical framework for developing programs that interact with data derived from human language in Python programming. Additionally, it includes over 50 corpora and lexical resources, such as WordNet, and a collection of text processing libraries for classification, tokenization, stemming, labeling, parsing, semantic reasoning, and wrappers for wrappers industrial-strength natural language processing (NLP) libraries. Porter stemming (or Porter stemmer) is a method for removing morphological and in flexional ends from English words. When setting information retrieval systems, the normalizing approach is most often used, and it is most commonly referred to as a "normalization process." In this work, the Porter stemming technique was used to extract morphological and in flexional ends from news articles, allowing for sentence normalization and mistake removal.

The TF-IDF technique is used to mitigate the effect of tokens that often occur in a dataset and are empirically less significant than features in a tiny fraction of the training dataset. The number of negative and positive words has been tallied in this study, with 0 positive and one negative word shown in Fig. 2.

The year-wise news and its positive and negative words have been measured in this work with Fig. 2 and X-axis denotes the fake news Subjects and Y-axis denotes the number of news counts.

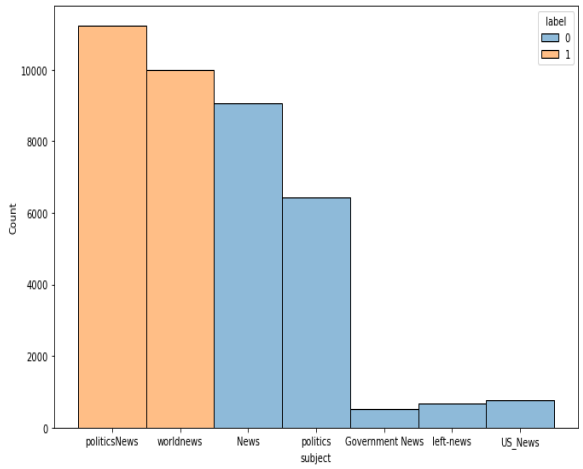


Fig. 2. Total Word Count.

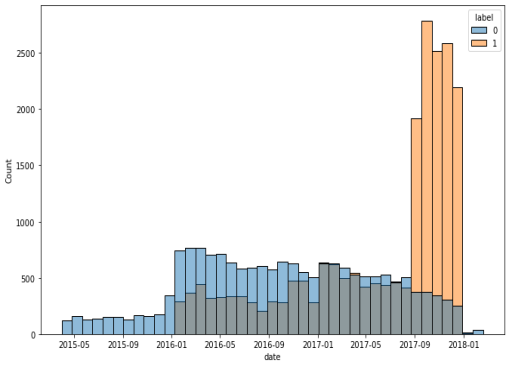


Fig. 3. Total Word Count on Date Wise.

Fig. 3 described that the greater number of FN had been measured in 2017 and 2019 and X-axis denotes the news feeding year and date and Y-axis denotes the number of news counts.

This study's word count and the TF-IDF calculated the average character length. The length of the false words character on the news content dataset was much longer than the length of the real words character due to the findings, as shown in Fig. 4 and X-axis denotes the Label like True or False and Y-axis denotes the Average character length.

Articles' propositions have been measured by counting the number of characters on each true text line and the number of characters on each erroneous text line in this work. The findings indicated that the fake character had used more articles from the news content dataset than the true character. The examination of propositions of articles on true and FN was depicted in Fig. 5, which represented the analysis results and X-axis denotes the number of characters and Y-axis denotes the proportion of articles.

In this work, the proposition of articles has also been measured in terms of the number of words per article on both the actual text and the false text, which has been considered. The findings revealed that the fake character had exploited more articles in the news content collection than the true character. The examination of propositions of articles on the number of words per piece of true and FN was depicted in Fig. 6, which showed the study results and X-axis denotes the number of words per article and Y-axis denotes the proportion of articles.

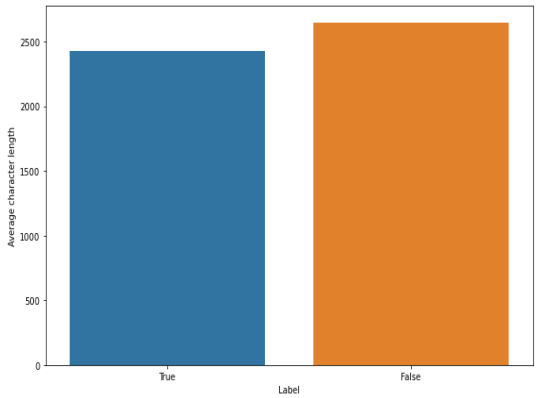


Fig. 4. Average Character Length.

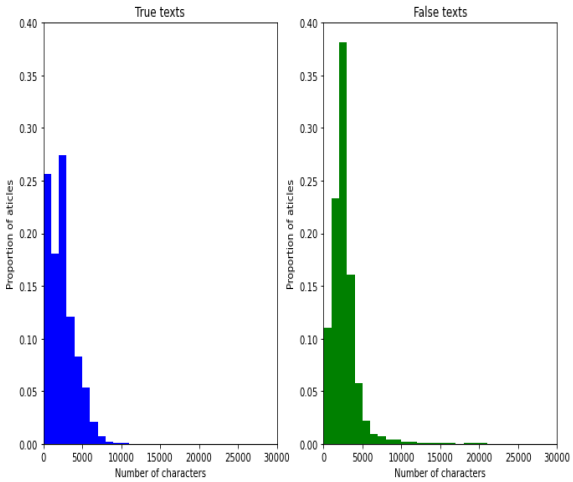


Fig. 5. Number of Character with Proposition of Article.

The work used a Fake URL data set to identify the most often encountered fake URL. The difference between fake and true news has been found in this effort to boost the learning rate of CNN. The sources of FN and authentic news were depicted in Fig. 7 and 8, respectively. With the source data set, 21 typical false news URLs were discovered, and the right-wing news website was shown to have distributed the fakest news based on the URL score and X-axis denotes the news count feed and Y-axis denotes the source of data.

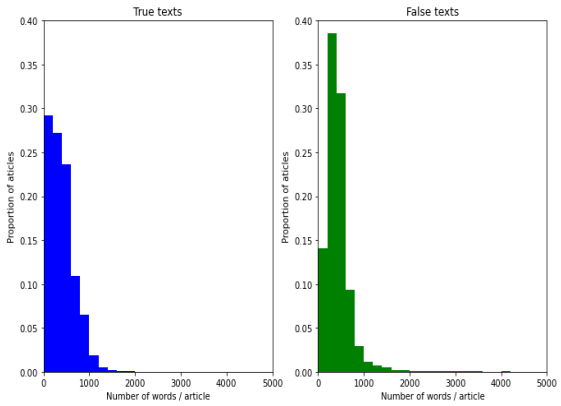


Fig. 6. Number of Character / Articles with Proposition of Article.

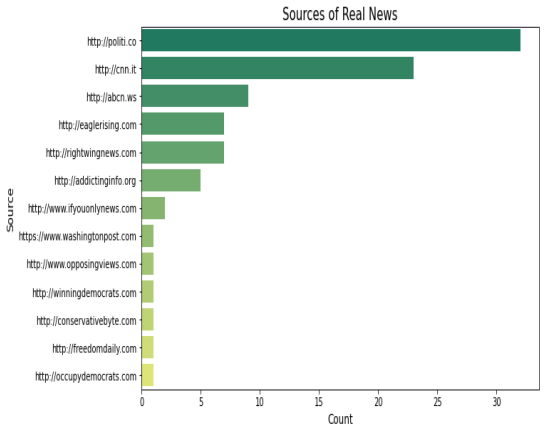


Fig. 7. Source of Real News.

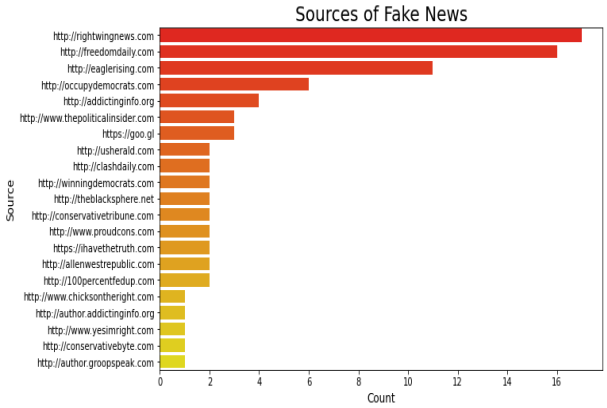


Fig. 8. Source of Fake News.

With this effort, Fig. 9 discovered a common source of actual and FN information and a common source of both. It has been determined that the seven most prevalent URLs are associated with this work and X-axis denotes the news count and Y-axis denotes the source of data.

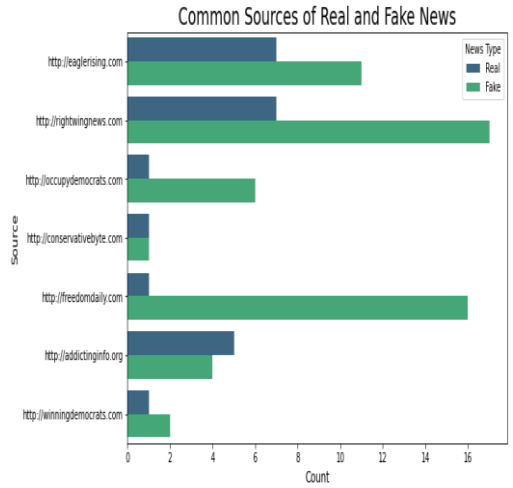


Fig. 9. Source of Common Real and Fake News.

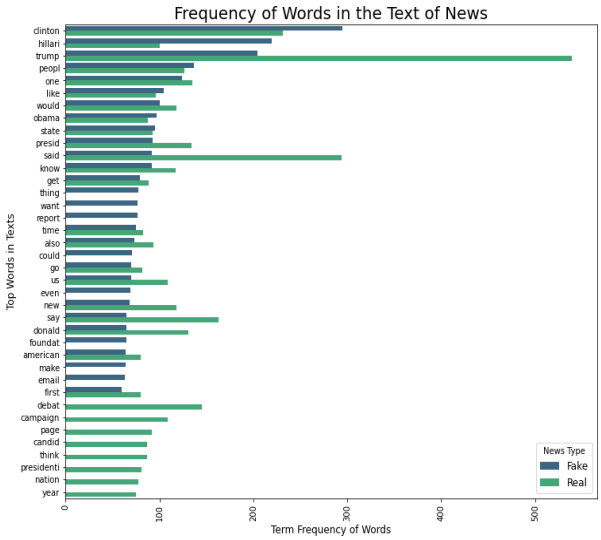


Fig. 10. Frequency Words in the Text of News.

A study was conducted to determine the frequency of word count in the news title and the news content to determine whether or not the news is fake. The chart depicted the frequency of news title appearances, whereas Fig. 10 indicated the frequency of text news appearances and X-axis denotes the Term Frequency of words and Y-axis denotes the Top words list.

Fig. 11 represents that the density of word distribution has been identified based on the frequency of words in the title and their distribution in the content by using the density parameter on the title. Among the findings was that the authentic news series has the largest dispersion density than the FN series. Using the graphic as an example, the outcome demonstrated the observations in the fictitious data are autocorrelated, which indicates that they are closely related to one another and the observations that came before them in the data set. The nature of the data makes it impossible to utilize typical FN detection algorithms to detect abnormalities in time series data due to the way the data is structured. To be successful, the task of time series false news identification must be handled independently of the other jobs. And X-axis denotes the title length and Y-axis denotes the density values.

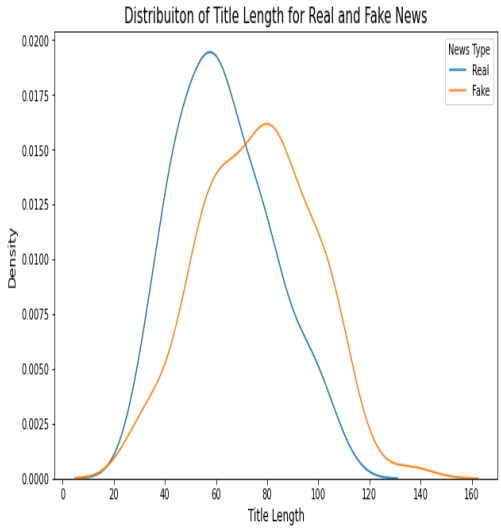


Fig. 11. Density Distribution Title Length for Real and Fake News.

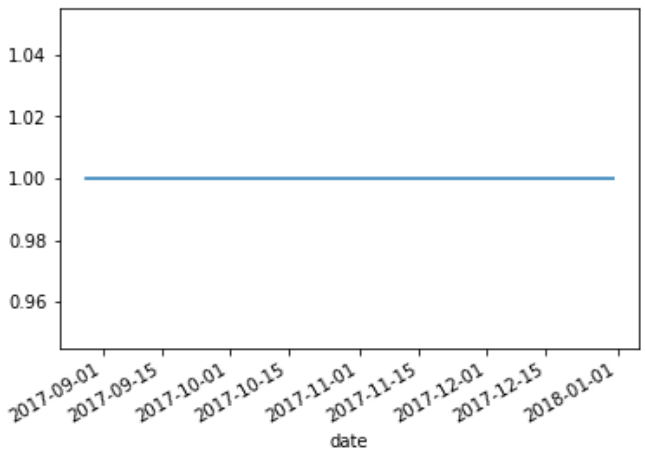


Fig. 12. Fake News Data Content Feed.

The count of news material has been autocorrelated with the time serious model developed by Arima in this research. Table II describes the autocorrelation on the FN data content, which depends on the feed's date, as depicted in Fig. 12 and X-axis denotes the date of fake news feed and Y-axis denotes the number of data. The result demonstrated that the news has been average with the date parameter. The autocorrelation of the lag that has been measured with this work has been signified in Fig. 13 and X-axis denotes the Lag Level and Y-axis denotes the Autocorrelation values. The density has also been calculated and exhibited in Fig. 14 and X-axis denotes the news content feed and Y-axis denotes the density percentage.

Table I represents the training and testing accuracy of Arima +Bi LSTM. The results show that the proposed Arima +Bi LSTM achieve a high accuracy of 0.9993. The performance has been evaluated in Fig. 15 and X-axis denotes the Epoch number and Y-axis denotes the accuracy percentage.

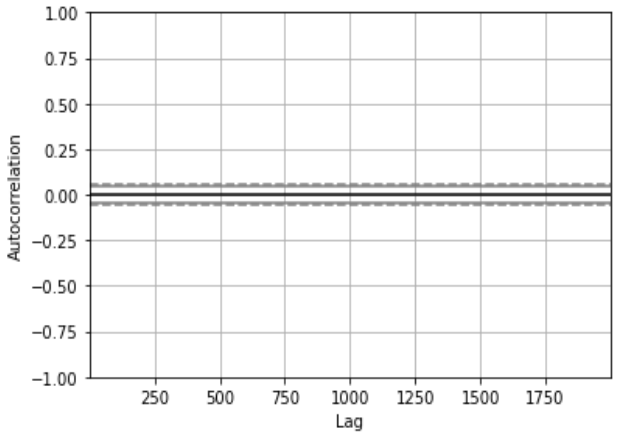


Fig. 13. Fake News Data Content Feed with Autocorrelation on Date.

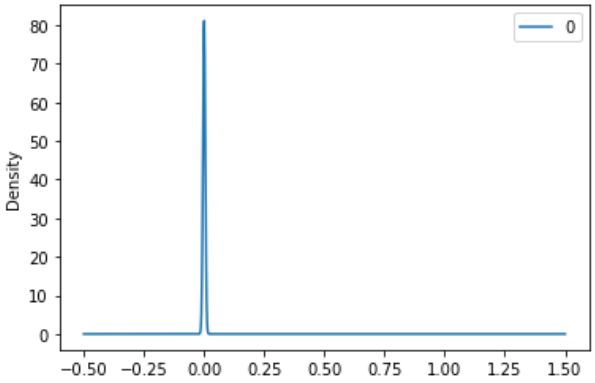


Fig. 14. Fake News Data Content Feed with Density Analysis.

TABLE I. ACCURACY ANALYSIS OF ARIMA+BI LSTM

Epoch	Training _Accuracy	Testing _Accuracy
1	0.8964	0.9937
2	0.9993	1.0000

Table II represents the training and testing loss of Arima +Bi LSTM. The results show that the proposed Arima +Bi LSTM achieve the minimal loss of 0.0032. The performance has been evaluated in Fig. 16 and X-axis denotes the Epoch number and Y-axis denotes the loss percentage.

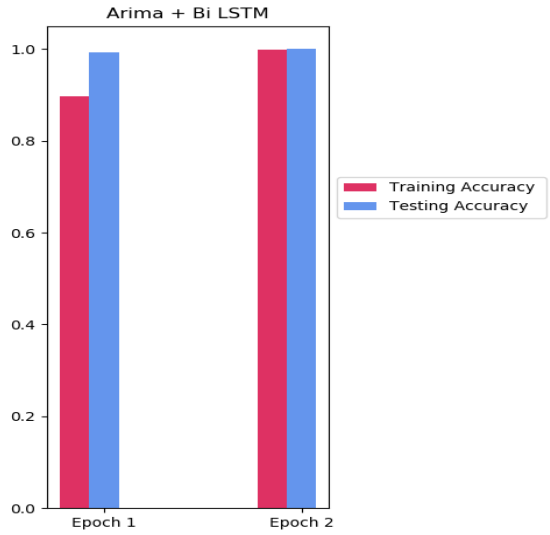


Fig. 15. Accuracy Analysis of Arima +Bi LSTM.

TABLE II. LOSS ANALYSIS OF ARIMA+BI LSTM

Epoch	Training _loss	Testing _Loss
1	0.2872	0.0163
2	0.0032	0.0014

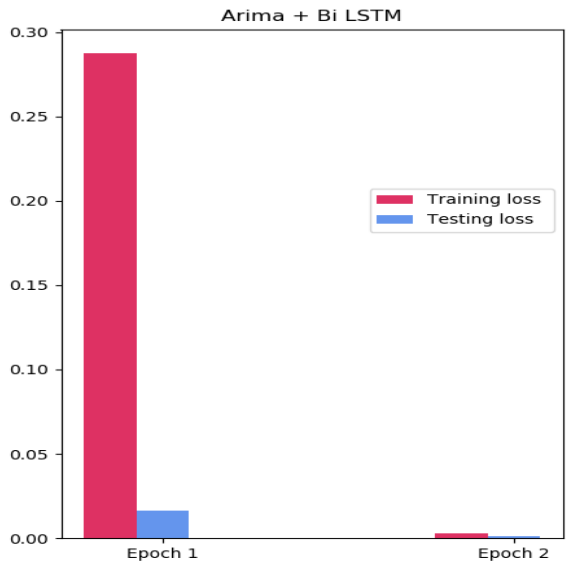


Fig. 16. Loss Analyses of Arima +Bi-LSTM.

Table III represents the training and testing accuracy of CNN. The results show that the CNN achieves an accuracy rate of 0.9676. The performance has been evaluated in Fig. 17 and X-axis denotes the Epoch numbers and Y-axis denotes the accuracy percentage.

TABLE III. ACCURACY ANALYSIS OF CNN

Epoch	Training _Accuracy	Testing _Accuracy
1	0.8502	0.9559
2	0.9319	0.9621
3	0.9450	0.9579
4	0.9592	0.9672
5	0.9676	0.9875

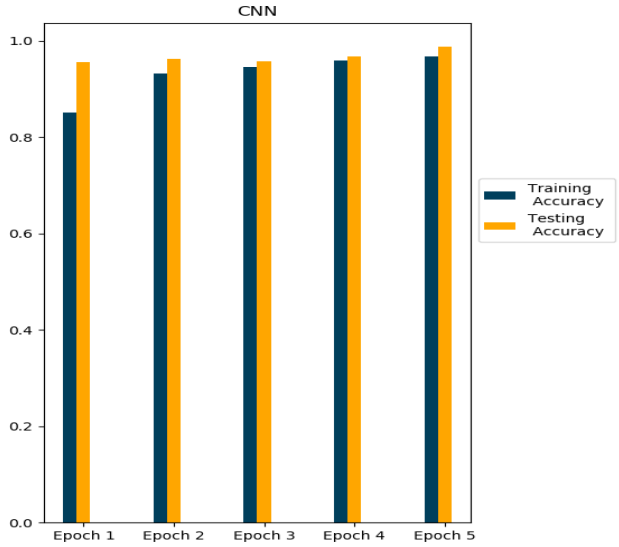


Fig. 17. Accuracy Analysis of CNN.

Table IV represents the training and testing Loss of CNN. The results show that the CNN achieves the loss rate of 0.1991. The performance has been evaluated in Fig. 18 and X-axis denotes the Epoch numbers and Y-axis denotes the loss values.

TABLE IV. LOSS ANALYSIS OF CNN

Epoch	Training _loss	Testing _Loss
1	2.8687	0.4365
2	0.3846	0.2573
3	0.2780	0.2928
4	0.2119	0.2012
5	0.1991	0.0630

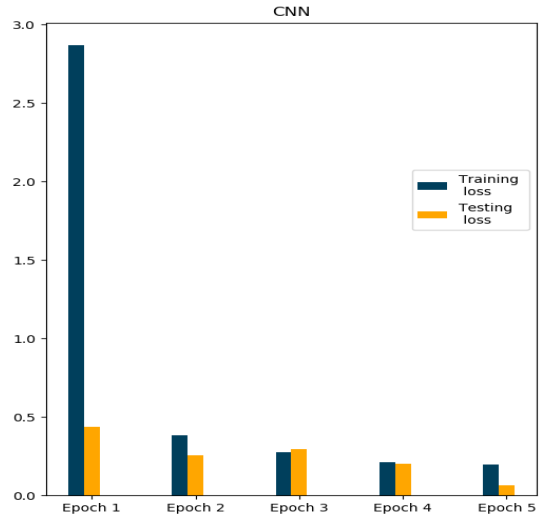


Fig. 18. Loss Analysis of CNN.

A. Evaluation Metrics

We employed many indicators to assess the performance of algorithms. The confusion matrix serves as the foundation for the majority of them. The confusion matrix is a tabular representation of the performance of a classification model on the test set, consisting of four parameters: true positive, false positive, true negative, and false negative.

1) Accuracy: Accuracy is a popular statistic representing the proportion of accurately anticipated observations, whether true or incorrect. The following equation may be used to determine the accuracy of model performance,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

In most circumstances, a model with a high accuracy value is a good model. Still, since we are training a classification model in this situation, an item that was predicted as true but was false (false positive) might have negative implications; similarly, an article forecasted as false but included factual data can generate trust concerns. Previously, we employed three different measures that considered the erroneously categorized observation, namely precision, recall, and F1-score.

2) Recall: Recall is a metric that indicates the total number of positive classifications outside the true class. Our example illustrates the proportion of articles expected to be true to the overall number of true articles and URLs.

$$Recall = \frac{TP}{TP+FN} \quad (17)$$

3) Precision: On the other hand, the accuracy score quantifies the ratio of true positives to all real occurrences anticipated. In our situation, precision refers to the proportion of articles classified as true among all positively predicted (true) articles,

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

4) F1-Score: The F1-score is a trade-off between accuracy and recall. It computes the harmonic mean of the two. thus, it considers both false positive and false negative observations. The following formula may be used to determine the F1-score,

$$F1 - Score = 2x\frac{Precision\ X\ Recall}{Precision+Recall} \quad (19)$$

Table V represents the performance metrics of Ensemble algorithms. The results show that the classification metrics are shown in Fig. 19. The DT, RF, SVM, and NB classifiers achieve 100% accuracy. X-axis denotes the Algorithms and Y-axis denotes the accuracy percentage.

Fig. 20 represents the FN detection using any new news content and Fig. 21 represent the FN detection with the URLs on the FLASK platform. Games, presentations, animations, visualizations, webpage components, and other interactive applications can all be made with Flask. The Python GUI platform was used for this project. A flashing indicator in Flask gives a very simple way to provide feedback to a user. This method allows you to record a message after each request and only access it on subsequent requests. This is often used in conjunction with a layout template that performs the same thing.

In Table VI is presented the Accuracy achieved in the existing authors and methods. And Fig. 22 represents the comparative accuracy chart for the existing author's method and proposed FNU-BiCNN method. In X-axis denotes the methods and Y-axis denotes accuracy percentage.

TABLE V. PERFORMANCE ANALYSIS OF ENSEMBLE MODEL

Methods	Accuracy	Precision	Recall	F-measure
DT	100	100	100	100
NB	100	100	100	100
RF	100	100	100	100
SVM	100	100	100	100
KNN	89	91	89	89
Ensemble (voting classifier)	100	100	100	100

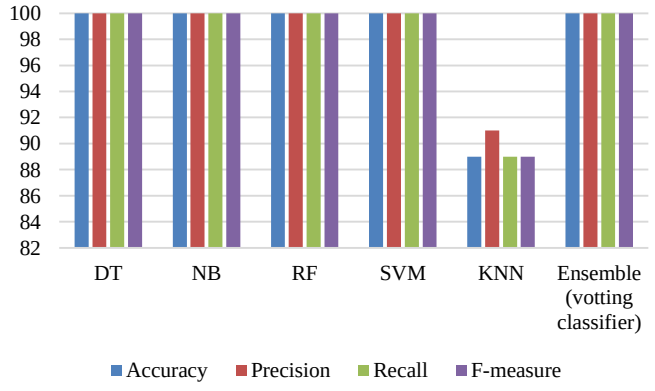


Fig. 19. Performance Analysis of Ensemble Model.

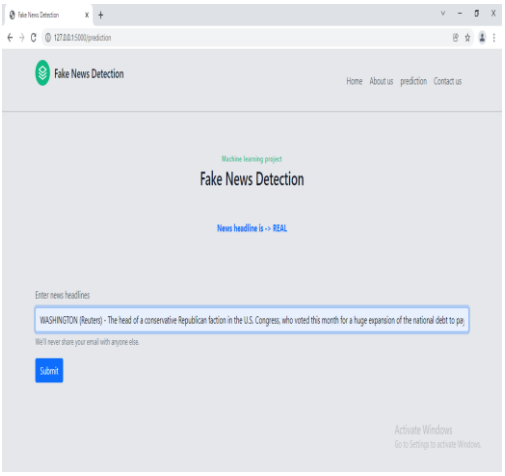


Fig. 20. Fake News Detection with FLASK Platform.

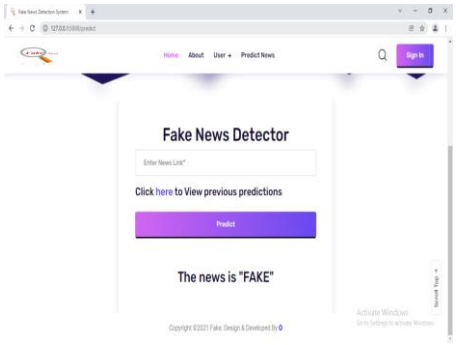


Fig. 21. Fake News Detection using URL Link.

TABLE VI. COMPARATIVE ANALYSIS

Paper number	Method	Accuracy
[1]	LSTM	97%
[4]	NB and RF	81%
[6]	NB	74%
[10]	CNN+LSTM	98.3%
[12]	DT	95%
[16]	LA-MSBD	95%
Proposed	FNU-BiCNN	100%

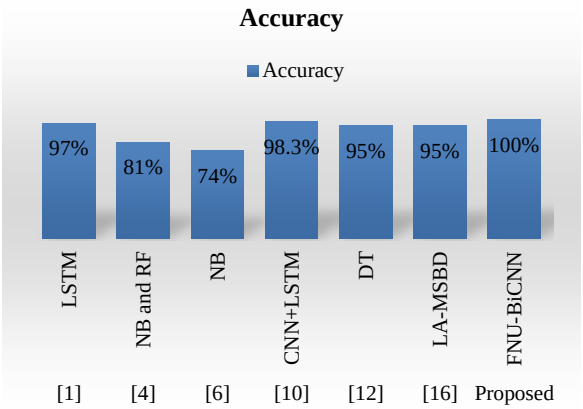


Fig. 22. Comparison Analysis for Various Authors and Proposed Method.

V. CONCLUSION

We attempted to evaluate fake and legitimate news in our study by thoroughly comprehending and fact-checking it. It outlines a broad methodology and the numerous aspects that affect the news's believability. Rather than studying a single strategy, we used an ensemble learning approach. The average accuracy score was 89% and 11% improvement over our worst-performing model, KNN. Additionally, we only utilized a portion of the information even with the supplied dataset and values. The FNU-BiCNN model is proposed in this article; Bi-LSTM was merged with an ARIMA model and CNN to reach the desired results, yielding a high accuracy rate of 0.9993 when combined with the other models. The proposed approach was evaluated using a variety of performance parameters, indicating that it yields an extraordinarily high accuracy rate of 100%. As the last step, our research creates an algorithm for identifying FN in the Kaggle dataset. Many messages and spam forwards confuse social network users by delivering inaccurate information. In our tests, the voting ensemble classifier came out on top regarding accuracy and precision above all other classification approaches.

Numerous outstanding difficulties with FN detection deserve the attention of researchers. For example, identifying key aspects involved in the propagation of FN is a critical first step toward reducing its spread. Similarly, real-time FN recognition in videos may be another future direction.

REFERENCES

[1] Agarwal, A., & Dixit, A. (2020). Fake News Detection: An Ensemble Learning Approach. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iciccs48265.2020.9121

[2] Ahmed, A. A., & Abdullah, N. A. (2016). Real-time detection of phishing websites. 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). doi:10.1109/iemcon.2016.7746247.

[3] Birunda, S. S., & Devi, R. K. (2021). A Novel Score-Based Multi-Source Fake News Detection using Gradient Boosting Algorithm. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). doi:10.1109/icais50930.2021.93958.

[4] Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake News Detection Using Sentiment Analysis. 2019 Twelfth International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2019.8844880.

[5] Cheng, W., Zhou, Y., Guo, Y., Hui, Z., & Cheng, W. (2019). Research on prediction method based on ARIMA-BP combination model. 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE). doi:10.1109/eitce47263.2019.90947.

[6] Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). doi:10.1109/ukrcon.2017.8100379.

[7] Hiramath, C. K., & Deshpande, G. C. (2019). Fake News Detection Using Deep Learning Techniques. 2019 1st International Conference on Advances in Information Technology (ICAIT). doi:10.1109/icait47043.2019.89872.

[8] Jiang, T., Li, J. P., Haq, A. U., & Saboor, A. (2020). Fake News Detection using Deep Recurrent Neural Networks. 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). doi:10.1109/iccwamtip51612.2020.9.

[9] Jayasiriwardene, T. D., & Ganegoda, G. U. (2020). Keyword extraction from Tweets using NLP tools for collecting relevant news. 2020 International Research Conference on Smart Computing and Systems Engineering (SCSE). doi:10.1109/scse49731.2020.931302.

[10] Kaliyar, R. K. (2018). Fake News Detection Using A Deep Neural Network. 2018 4th International Conference on Computing Communication and Automation (ICCCA). doi:10.1109/ccaa.2018.8777343.

[11] Konkobo, P. M., Zhang, R., Huang, S., Minoungou, T. T., Ouedraogo, J. A., & Li, L. (2020). A Deep Learning Model for Early Detection of Fake News on Social Media*. 2020 7th International Conference on Behavioural and Social Computing (BESC). doi:10.1109/besc51023.2020.934831.

[12] Lyu, S., & Lo, D. C.-T. (2020). Fake News Detection by Decision Tree. 2020 SoutheastCon. doi:10.1109/southeastcon44009.202.

[13] Mansouri, R., Naderan-Tahan, M., & Rashti, M. J. (2020). A Semi-supervised Learning Method for Fake News Detection in Social Media. 2020 28th Iranian Conference on Electrical Engineering (ICEE). doi:10.1109/icee50131.2020.9261053.

[14] Masood, F., Ammad, G., Almogren, A., Abbas, A., Khattak, H. A., Din, I. U., … Zuair, M. (2019). Spammer Detection and Fake User Identification on Social Networks. IEEE Access, 1–1. doi:10.1109/access.2019.2918196.

[15] Qazi, M., Khan, M. U. S., & Ali, M. (2020). Detection of Fake News Using Transformer Model. 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). doi:10.1109/icomet48670.2020.9074.

[16] Rout, R. R., Lingam, G., & Somayajulu, D. V. L. N. (2020). Detection of Malicious Social Bots Using Learning Automata With URL Features in Twitter Network. IEEE Transactions on Computational Social Systems, 1–15. doi:10.1109/tcss.2020.2992223.

[17] Shantanu, Janet, B., & Joshua Arul Kumar, R. (2021). Malicious URL Detection: A Comparative Study. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). doi:10.1109/icais50930.2021.9396.

[18] Singh, L. (2020). Fake News Detection: a comparison between available Deep Learning techniques in vector space. 2020 IEEE 4th Conference on Information & Communication Technology (CICT). doi:10.1109/cict51604.2020.931209.

[19] Seo, Y., & Jeong, C.-S. (2018). FaGoN: Fake News Detection model using Grammatic Transformation on Neural Network. 2018 Thirteenth International Conference on Knowledge, Information and Creativity Support Systems (KICSS). doi:10.1109/kicss45055.2018.89505.

[20] Vogel, I., & Meghana, M. (2020). Detecting Fake News Spreaders on Twitter from a Multilingual Perspective. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). doi:10.1109/dsaa49011.2020.00084.

[21] Zhi, X., Xue, L., Zhi, W., Li, Z., Zhao, B., Wang, Y., & Shen, Z. (2021). Financial Fake News Detection with Multi fact CNN-LSTM Model. 2021 IEEE 4th International Conference on Electronics Technology (ICET). doi:10.1109/icet51757.2021.945092.

Dynamic Vehicular Communication using Gaussian Interpolation of Cluster Head Selection (GI-CHS)

Mahmoud Zaki Iskandarani
Faculty of Engineering
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—Decentralized and centralized vehicular communication is investigated in this work using Gaussian interpolation function with cluster head (CH) selection technique. The work uncovered that the best communication approach is to use both centralized and decentralized vehicular communication as combining them will achieve a much more uniform results as a function of communication radius values and vehicular speed. It is also found that vehicular speed contributes negatively to the efficiency of data communication if the relative speed of the vehicles to the communication radius is limited by their ratios. Mathematical expression is presented that relates probability of successful transmission to communication radius for both centralized and decentralized techniques with data proving the importance of the spread parameter within the Gaussian interpolation in a tabulated form, and explained to prove the adaptability of the function used. It is also shown in this work that weights affecting CH selection, thus using Gaussian interpolation is proved to be important as a weighting function in an a adaptive and dynamic vehicular ad-hoc networks (VANETS) covering both vehicle to vehicle (V2V), and vehicle to infrastructure (V2I) communication through cluster head selection.**

Keywords*—Cluster head; VANETS; adaptive routing; weighted clustering; Gaussian interpolation; V2V; V2I

I. Introduction

Due to the increase in connected and autonomous vehicles and number of vehicles in urban areas, Vehicular Ad-hoc Networks (VANETs) clustering has become increasingly crucial. For drivers and passengers, VANETs can provide safety-related apps, Internet connectivity, and a variety of user applications.

The real-time identification of road conditions as a function of fast vehicle movement and topological changes, which necessitates the development of dynamic routing protocols, is a difficult issue for VANETs. By increasing connectivity times with better signal quality and improving routing performance due to scalability, clustering can significantly contribute to more efficient bandwidth utilization, dependable message exchange and delivery. As a result of grouping traveling vehicles into clusters, network performance can be greatly improved [1], [2], [3].

Clustering is a process of grouping vehicles regarded as nodes of a network into groups forming hierarchical structure. This structure provides specific functions leading to better quality of service (QoS). Consequently, neighboring nodes representing vehicles can join a cluster according to stated metrics. Generally, the cluster structure has three main types of nodes (vehicles): a Cluster Head (CH), Cluster Member (CM), and Cluster Gateway (CG). Traditionally, a way to choose a CH is to regard the first vehicle moving in a certain direction as CH, then Vehicles, within the predefined parameters to CH are grouped together, thus forming a multi-hop cluster. However, recently instead of just choosing the first vehicle as a CH, clustering mechanisms are calculated based on efficiency and stability of a vehicle (node) to its surrounding environment [4], [5], [6], [7], [8], [9], [10].

Clustering approach supports direct interaction between clusters of vehicles, which VANET routing protocols use to improve traffic efficiency and achieve traffic optimization and increase mobility and road safety through Vehicle to Vehicle communication under cooperative driving principles. In VANETs, vehicles have onboard sensing systems and transceivers that facilitate V2V communication directly, which allows real time exchange of important information with low latency independent of road side infrastructure. However, under certain conditions, V2V communication requires road side infrastructure to enable other safety, mobility and environmental supply of data to the travelling vehicles, provided through road side units (RSUs), thus forming vehicle to infrastructure communication (V2I) [11], [12], [13], [14], [15].

VANETs covers geographically varying networks with changing dynamic range and mobility; so it is essential to have efficient routing for VANET environments. Clustering is one among the major classification in energy efficient Vehicular Ad Hoc Networks (VANETs covering vehicle to vehicle and vehicle to infrastructure communication. Clustering is classified according to vehicle position, traffic density and congestion level as a function of mobility [16], [17], [18], [19], [20].

II. Background

VANETs suffer from variables affecting their communication and data exchange process such as:

1) High latency
2) Data security
3) Routing and routes length
4) Channel congestion

To enable quality of service and better resource management, hierarchical structure is proposed by researchers [21], [22]. Such approach describes the process of close to each other vehicles, with shared features, to join a group that is termed a cluster. In VANETs, the process of clustering involves a cluster head (CH), which has the main task in the formation process of a cluster. Such a cluster can be formed in different ways according to selected metrics.

CH can be considered as a mobile routing node with CM that represent a vehicle as a standard mobile node, where CG can be formed by two CMs with an interfacing task. Researchers considered metrics to select CH and a cluster such as:

1) Average vehicular velocity
2) Average vehicular acceleration
3) Vehicular position
4) Vehicular heading
5) Traffic Density

Coherence, continuity and stability with consideration of the above parameters are critical in CH is selection and subsequent membership associated with each selected CH. Optimizing issues such as routing within a cluster (intra-cluster) and between CHs and CHs (inter-cluster) are dependent on all the previous parameters consideration [23].

Research is based on the static approach that depends on cluster forming based vehicles closeness to the Base Station and or Road Side Unit in order to choose Cluster Head. Other research work focused on dynamic and adaptive clustering, which eliminates cluster formation based on RSU and considers the other metrics such as speed of vehicles, destination, and movement pattern in order to form a cluster [24], [25].

Thus, clustering is critical element in VANET routing protocols, whereby a group of nodes is organized to form a temporary network on the road on the basis predefined metrics. This approach makes the network more reliable and controllable as a cluster head (CH) is selected for vehicular group on the basis of defined parameters, with the rest of the nodes (vehicles) become cluster members (CMs). The chosen CH takes on the responsibility for managing the CMs and for intra-cluster communication, which reduces Basic Safety Messages (BSMs) delivery times. A good clustering approach for selecting a CH is to choose the member with highest metrics in terms of ability to lead the temporary network for the longest time interval with high storage and processing capacities [26], [27], [28].

In this paper analysis of vehicular communication is carried out using CHs approach and Gaussian interpolation function, which is employed to enable adaptive clustering by using both communication radius and vehicular velocity in order to provide more efficient communication and reliable routes. The work differs from previous research by incorporating of combined Gaussian interpolation and ratio functions into one function covering communication variation in two dimensions (distance, velocity).

The rest of this paper is divided as follows: Background Methodology, Results and Discussion, Conclusions, References.

III. Methodology

VANET is essential for both safety based message exchanges for vehicles. Thus, it is important that an optimum routing algorithm is achieved with clustering taken into consideration, to enable efficient and effective V2V and V2I data transfer. By applying message exchange techniques to groups of vehicles (clusters) with Gaussian interpolation function, more efficient communication channel utilization can be achieved.

Due to the dynamic nature of VANETs, V2V and V2I communication could bear some data loss resulting from connectivity interruption as a result of vehicular (nodes) movement. Thus, it is critical after selecting CH to continuously update nodes positions (trajectories). To help this process, zones can be created per area under consideration with Gaussian interpolation function used to enable smooth and continuous transmission and data exchange among CHs, thus acting as bridging or linking CHs (BCH). So, CHs will coordinate communications between cluster' members under certain criteria with each selected CH communicates messages to vehicles known as cluster members (CMs). This approach of CH, CM and BCH, reduces routing cost and delay that could result during data exchange.

As a result of the dynamic and mobile nature of vehicular communication, there is a need for an adaptive approach to vehicular clustering. The approach in this work is based on:

1) Utilization of Gaussian interpolation as bridge for cluster heads (BCHs).
2) Application of zoning to enable smother, less congested communication, with better management.
3) Implementation of both Communication Radius and Vehicular Speed in combination with the Gaussian interpolation function.
4) To enable characterization of the benefits using Gaussian interpolation function, standard approach with multi hop routing is simulated using:

a) Equally weighted CH position (Communication Radius) and vehicular Speed.

b) Gaussian weighted CH position (communication radius) and vehicular speed.

To carry simulation for the two approaches, zoning is used as a first step in order to better analyze the outcomes. Equation (1) show the zoning expression.

$$Zones_{(N-lanes)} = \left(\frac{Lane\ Length}{Zone\ Length}\right) * N \quad (1)$$

The simulation area is divided into 6 zones , each zone width is a 100 meters wide. Within each zone and along the travelled path, vehicles (nodes) will form clusters and VANET clouds and exchange data (Basic Safety Messages). Choosing cluster head and cluster members is carried.

Equation (2) show the implementation of Gaussian interpolation function used to compute CH selection weight CH_w with reachability parameter (δ) [29], [30].

$$CH_w = \sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{\left(\left(\frac{Location_{CH}}{R_{Comm}}\right)-1\right)^2}{2*\delta}\right)\right) \quad (2)$$

Equation (2) can be simplified and result in equation (3).

$$CH_w = \sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{(\theta-1)^2}{2*\delta}\right)\right) \quad (3)$$

Examining equation (3), three conditions are applied as follows:

$$\theta \to 1 \Rightarrow CH_{weight} \to 1 \quad (4)$$

$$\theta \to 0 \Rightarrow CH_w = \sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{-1}{2*\delta}\right)\right) \quad (5)$$

If vehicular movements are outside the effective communication radius, then equation (4) will be reduced to equation (6).

$$\theta \gg 0 \Rightarrow CH_w = \sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{\theta^2}{2*\delta}\right)\right) \quad (6)$$

The previous equations show the adaptive behavior of clustering and cluster head selection using Gaussian interpolation function, when computing weights in a dynamic environment, such as vehicular movements. This is not present in the standard fixed weight approach for computing effective CH selection, which with the membership of nodes will definitely affect efficiency of data exchanged, energy consumed, routes travelled and transmission delays.

IV. Results and Discussion

Fig. 1 shows the relationship between probability of successful transmission (P_s) and Communication Radius using Centralized (V2I) communication. The transmission is within a maximum radius of 200 meters (The width of two lanes) specified in the simulation. It is clear that as the communication radius increases, so does the efficiency of data transmission due to the following:

1) Increase of the transmission range.
2) Increase in the number of RSUs along and across the road.
3) Inclusion of more RSUs and Vehicles.

Thus errors and probability of error decreases as the communication radius increases. It is noticeable that at low radius values and due to distance between vehicles and road side units (RSUs) the efficiency of data delivery is very low, which results in higher probability of transmission error and data loss.

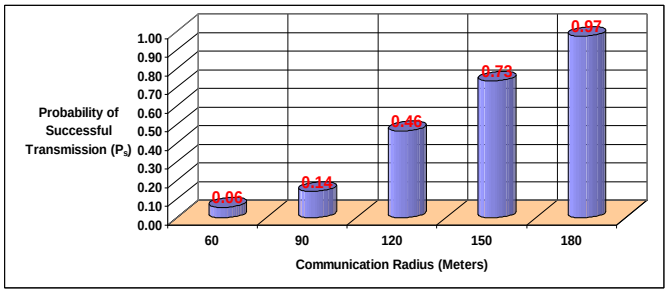


Fig. 1. Centralized Transmission Efficiency.

Fig. 2 shows the relationship between probability of successful transmission (P_s) and Communication Radius using decentralized and dynamic (V2V) communication. The transmission is within a maximum radius of 200 meters (The width of two lanes) specified in the simulation. It is clear that as the communication radius increases, so does the efficiency of data transmission due to the following:

1) Increase of the transmission range.
2) Increase in the number of vehicles, thus increase in the number of dynamic cluster heads.
3) Inclusion of more Vehicles, thus more cluster members (CMs).
4) The dynamic interaction between CMs and CHs and among CHs.

Thus errors and probability of error decreases as the communication radius increases. There is a clear difference between decentralized and centralized transmission efficiency characteristics. In the decentralized response in Fig. 2, a smoother response with higher levels of efficiency for small radius values compared to centralized communication.

Effect of combining both centralized and decentralized approaches in communication is shown in Fig. 3. An obvious improvement for small radius values efficiency due to the dynamic movement and dynamic communication (V2V) that covered some of the shortcomings of (V2I) communication. Also, clear, smooth and gradual increase in efficiency as a function of radius values for a fixed Gaussian spread value of δ=5 is witnessed [31]. Thus despite the small drop in efficiency at high distances, the overall response is much more favorable, and reliable the either one used independently.

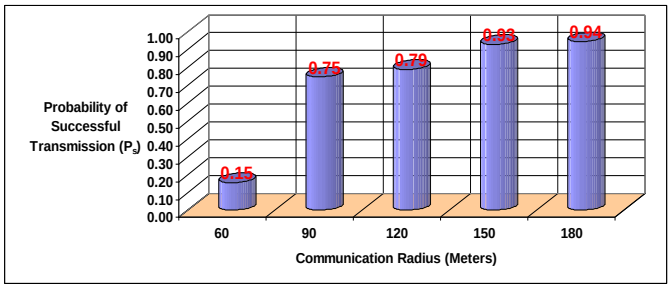


Fig. 2. Decentralized Transmission Efficiency.

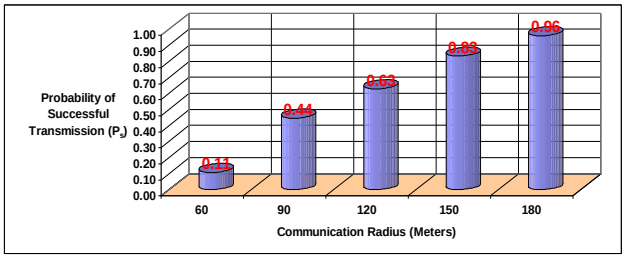


Fig. 3. Combined Transmission Efficiency.

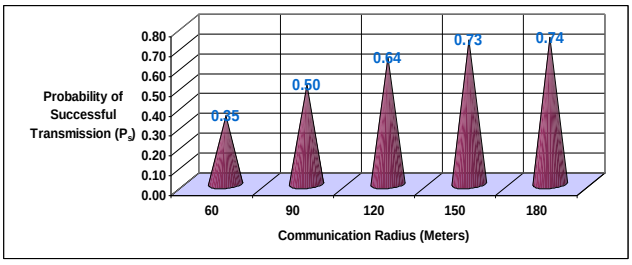


Fig. 4. Combined Transmission Efficiency with Vehicular Speed Effect

Due to vehicular movement, the probability of successful transmission (P_s) can be affected by vehicular speed depending the speed of the vehicle. Fig. 4 show effect of vehicular speed on probability of successful transmission per traveled communication radius. The figure clearly shows that as the travelled distance decreases relative to vehicular speed, the communication efficiency is reduced as well. Thus, better efficiency is obtained when either the speed is slow within a narrow radius or when the speed is either slow or fast within a large communication radius.

Combining decentralized and centralized vehicular communication can be considered an optimum solution to vehicular communication. This is particularly true since simulation showed that when the communication radius value is low, the efficiency of centralized communication drops dramatically with very high reduction in the probability of successful transmission. This is shown in Fig. 5.

As Fig. 5 shows the efficiency of decentralized communication for small communication coverage areas is 10 times higher in the case of decentralized (V2V) communication, which is due to the coverage area. Also, the overall probability of successful transmission drops for both centralized and decentralized communication as number of vehicles is reduced and also due to relative increase of vehicular speed as coverage area is reduced. It is worth mentioning that increase in vehicular speed contributes to the reduction of the overall available vehicles (CMs) per considered area, which leads to drop in communication efficiency.

Initial expression that relates centralized to decentralized efficiency is given in equation (7)

$$P_s(Decentralized) = \psi * P_s(Centralized) \quad (7)$$

Where:

ψ: Optimizing parameter related to the relative vehicular speed to communication radius.

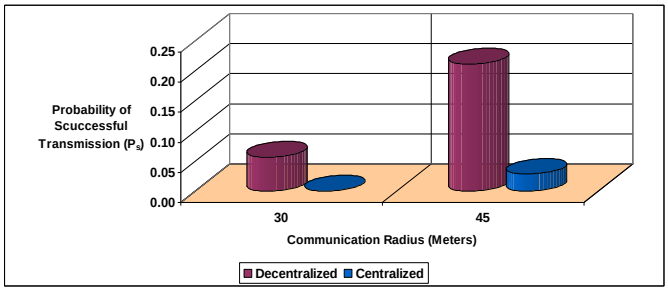


Fig. 5. Transmission Efficiency at Low Radius Values.

The plot in Fig. 5 supports the finding from Fig. 1 and 2, which show that centralized communication is more efficient at wider radius coverage compared to the decentralized one, which is affected by the dynamics of vehicular movement. Thus, the figure further proof that centralized communication has lower efficiency for low radius values. What Fig. 5 shows in addition to that is the usefulness of decentralized communication at low communication radius values, where centralized efficiency falls to near zero. This helps the centralized communication when combined with the decentralized one.

From the simulated data and equation 2, a relationship can be shown between probability of successful transmission and the Gaussian weight function that relates radius to travelled distance and to speed as shown in equations (8).

$$P_s(Combined) = \kappa * \left(\sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{\left(\left(\frac{Location_{CH}}{R_{Comm}}\right)-1\right)^2}{2*\delta} \right) \right) \right)^{\phi} \quad (8)$$

Where:

κ: Multiplication coefficient ($1 \leq \kappa \leq 2$)

ϕ: Power coefficient related to the relative vehicular speed to communication radius ($2 \leq \phi \leq 6$).

Using equation (3), equation (8) is represented as in equation (9).

$$P_s(Combined) = \kappa * \left(\sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{(\theta-1)^2}{2*\delta} \right) \right) \right)^{\phi} \quad (9)$$

From equations (7) and (9), expressions covering centralized and decentralized communication are obtained and shown in equations (10) and (11).

$$P_s(Centralized) = \left(\frac{2}{1+\psi}\right) * \left(\kappa * \left(\sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{(\theta-1)^2}{2*\delta} \right) \right) \right)^{\phi} \right) \quad (10)$$

$$P_s(Decentralized) = \left(\frac{2*\psi}{1+\psi}\right) * \left(\kappa * \left(\sum_{\substack{R_{Comm}=R_1 \\ \delta=1}}^{\substack{R_{Comm}=R_K \\ \delta=Q}} exp\left(-\left(\frac{(\theta-1)^2}{2*\delta} \right) \right) \right)^{\phi} \right) \quad (11)$$

The expressions in equations (10) and (11) can be further simplified as shown in equations (12) and (13).

$$P_s(Centralized) = \left(\frac{2}{1+\psi}\right) * \Omega \quad (12)$$

$$P_s(Decentralized) = \left(\frac{2*\psi}{1+\psi}\right) * \Omega \quad (13)$$

From equations (12) and (13) and assuming that $\psi \gg 1$, equations (14) and (15) are obtained.

$$P_s(Centralized) = \left(\frac{2*\Omega}{\psi}\right) \quad (14)$$

$$P_s(Decentralized) = 2 * \Omega \quad (15)$$

When $\psi \ll 1$, equations (12) and (13) are reduced to equations (16) and (17).

$$P_s(Centralized) = 2 * \Omega \quad (16)$$

$$P_s(Decentralized) = 2 * \psi * \Omega \quad (17)$$

The expressions in equations (14) to (17) proof that the centralized process alone is less efficient than the decentralized and a balanced solution would be to combine both techniques. The equations also show an interesting result, whereby the centralized approach and decentralized approach converge to the same expression but at opposite sides of ψ. This is further proof that they can be employed as complementary techniques.

Table I present example of weights generated used Gaussian interpolation function. The tabulated weights show that as δ increases, the reachability of the function increases per fixed speed, thus referring to equation (8), will result in increase in the communication efficiency. Also, as the communication radius increase, the reachability of the interpolation expands to cover such increase, thus a more uniform communication and data exchange occurs with dynamic response to communication radius increment.

TABLE I. EXAMPLE OF GAUSSIAN COMPUTED WEIGHTS USED FOR CH SELECTION

Radius	***R1***	***R2***	***R3***	***R4***	***R5***	
	60	***90***	***120***	***150***	***180***	
Gaussian Interpolation	CH_{w1}	CH_{w2}	CH_{w3}	CH_{w4}	CH_{w5}	***δ***
	0.07	**0.23**	**0.35**	**0.43**	**0.47**	**1**
	0.26	**0.48**	**0.59**	**0.65**	**0.69**	**2**
	0.40	**0.61**	**0.70**	**0.75**	**0.78**	**3**
	0.51	**0.69**	**0.77**	**0.81**	**0.83**	**4**
	0.58	**0.74**	**0.81**	**0.84**	**0.86**	**5**

V. CONCLUSION

In this work, a simulation based investigation covering centralized and decentralized vehicular communication is carried out successfully. The simulated work analyzed the probability of successful transmission, which is indicative of communication network reliability and efficiency using Gaussian interpolation function as an adaptive weight function with integrated ratio component.

The presented work focused on three main elements:

1) Proving that combining centralized and decentralized clustering communication process is a viable answer to obtain balanced process.

2) Employing Gaussian interpolation with weight ratio as a smooth transition function that is able to avoid abrupt communication quality changes, thus improving QoS.

3) Establishing that a combination of vehicular velocity and communication distance as the fundamental parameters that have main effect on transmission efficiency with emphasis on the problem of vehicular velocity that has marked effect on efficiency, specifically at high speeds in relation to communication distance, which also affects cluster size.

The weight function is used to enable more efficient clustering, particularly when selecting cluster heads (CHs). The obtained data at different communication distances proved that decentralized communication has more uniform connectivity, especially at low communication distances, while centralized communication has higher efficiency at larger distance as it does not suffer from some of the dynamics that decentralized communication faces.

It is also shown that the spread δ used in the Gaussian interpolation affect probability of successful transmission. Mathematical models describing the relationship between the Gaussian interpolation function and transmission efficiency showed that this presented model and technique can be optimized through three parameter (δ, ψ, ϕ). Vehicular speed is found to reduce attained efficiency due to the dynamic relation between radius values and vehicular speed, which affects signal stability.

Further work is needed in terms of establishing a more optimized CH selection criteria using Gaussian interpolation with the variable θ assigned other functions taking into account rural and urban areas.

ACKNOWLEDGMENT

I would like to acknowledge previous MATLAB simulation work carried out by Mr. Amburose Sekar.S.

REFERENCES

[1] P. Sondi, I. Abbassi, E. Ramat, E. Chebbi, and M. Graiet, "Modeling and verifying clustering properties in a vehicular ad hoc network protocol with Event-B," Sci. Rep., vol. 11, no. 1, pp. 1–15, 2021, doi: 10.1038/s41598-021-97063-3.

[2] R. Kaur, R. K. Ramachandran, R. Doss, and L. Pan, "The importance of selecting clustering parameters in VANETs: A survey," Comput. Sci. Rev., vol. 40, p. 100392, 2021, doi: 10.1016/j.cosrev.2021.100392.

[3] S. Kad and V. K. Banga, "A systematic classification of routing in vehicular ad hoc networks," Int. J. Eng. Adv. Technol., vol. 8, no. 6, pp. 5336–5355, 2019, doi: 10.35940/ijeat.F8779.088619.

[4] O. Senouci, S. Harous, and Z. Aliouat, "Survey on vehicular ad hoc networks clustering algorithms: Overview, taxonomy, challenges, and open research issues," Int. J. Commun. Syst., vol. 33, no. 11, pp. 1–21, 2020, doi: 10.1002/dac.4402.

[5] M. Ren, J. Zhang, L. Khoukhi, H. Labiod, and V. Veque, "A Unified Framework of Clustering Approach in Vehicular Ad Hoc Networks," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 5, pp. 1401–1414, 2018, doi: 10.1109/TITS.2017.2727226.

[6] J. RaoDawande, S. Silakari, and A. Deen, "A Survey of all Existing Clustering Protocols in VANETS but Main Emphasis of Survey Laid on Currently using Protocol i. e TCDGP," Int. J. Comput. Appl., vol. 118, no. 6, pp. 22–31, 2015, doi: 10.5120/20751-3146.

[7] F. Yang and Y. Tang, "Cooperative clustering-based medium access control for broadcasting in vehicular ad-hoc networks," IET Commun., vol. 8, no. 17, pp. 3136–3144, 2014, doi: 10.1049/iet-com.2014.0397.

[8] H. Wang, R. P. Liu, W. Ni, W. Chen, and I. B. Collings, "VANET modeling and clustering design under practical traffic, channel and mobility conditions," IEEE Trans. Commun., vol. 63, no. 3, pp. 870–881, 2015, doi: 10.1109/TCOMM.2015.2388575.

[9] S. Ucar, S. C. Ergen, and O. Ozkasap, "Multi-Hop Cluster based IEEE 802. 11p and LTE Hybrid Architecture for VANET Safety Message Dissemination," no. 11315, 2013.

[10] A. F. M. S. Shah, M. A. Karabulut, H. Ilhan, and U. Tureli, "Performance optimization of cluster-based MAC protocol for VANETs," IEEE Access, vol. 8, pp. 167731–167738, 2020, doi: 10.1109/ACCESS.2020.3023642.

[11] R. S. Tomar, S. Verma, and G. S. Tomar, "Cluster based RSU centric channel access for VANETs," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7420, pp. 150–171, 2013, doi: 10.1007/978-3-642-35840-1_8.

[12] R. Adrian, S. Sulistyo, I. W. Mustika, and S. Alam, "A controllable rsu and vampire moth to support the cluster stability in vanet," Int. J. Comput. Networks Commun., vol. 13, no. 3, pp. 79–95, 2021, doi: 10.5121/ijcnc.2021.13305.

[13] M. Ben Bezziane, A. Korichi, C. A. Kerrache, and M. E. A. Fekair, "Rcvc: Rsu-aided cluster-based vehicular clouds architecture for urban areas," Electron., vol. 10, no. 2, pp. 1–18, 2021, doi: 10.3390/electronics10020193.

[14] T. Omar, K. Guerra, C. Mardoyan, S. Sharma, and X. Rangel, "Smart Cities V2I Cloud based Infrastructure using Road Side Units," no. IoTBDS, pp. 270–277, 2021, doi: 10.5220/0010469402700277.

[15] D. Roy, "Trust and Group Leader based Model to Avoid Broadcast Storm Problem in Vehicular Ad-hoc Networks," vol. 10, no. 4, pp. 575–597, 2017.

[16] S. A. Rashid, L. Audah, M. M. Hamdi, and S. Alani, "Prediction based efficient multi-hop clustering approach with adaptive relay node selection for VANET," J. Commun., vol. 15, no. 4, pp. 332–344, 2020, doi: 10.12720/jcm.15.4.332-344.

[17] M. Jalasri and L. Lakshmanan, "Code-based encryption techniques with distributed cluster head and energy consumption routing protocol," Complex Intell. Syst., no. 0123456789, 2021, doi: 10.1007/s40747-021-00505-8.

[18] A. Temurnikar, P. Verma, and G. Dhiman, "A PSO enable multi-hop clustering algorithm for VANET," Int. J. Swarm Intell. Res., vol. 13, no. 2, pp. 1–14, 2022, doi: 10.4018/IJSIR.20220401.oa7.

[19] M. Najafi and M. R. S. Aghaei, "An Efficient Cluster-based Routing Protocol for Improvement Delay in Mobile Ad-hoc Networks," no. January, 2020, [Online]. Available: http://dx.doi.org/10.20944/preprints202001.0342.v2.

[20] S. David and P. T. Vanathi, "Middle-Order Vehicle-Based Clustering Model for Reducing Packet Loss in Vehicular Ad-hoc Networks," J. Circuits, Syst. Comput., vol. 29, no. 11, pp. 1–16, 2020, doi: 10.1142/S0218126620501807.

[21] S. Ebadinezhad, "Design and performance evaluation of Improved DFACO protocol based on dynamic clustering in VANETs," SN Appl. Sci., vol. 3, no. 4, pp. 1–15, 2021, doi: 10.1007/s42452-021-04494-8.

[22] M. A. Saleem et al., "Deep Learning-Based Dynamic Stable Cluster Head Selection in VANET," J. Adv. Transp., vol. 2021, 2021, doi: 10.1155/2021/9936299.

[23] B. Elira, K. P. Keerthana, and K. Balaji, "Clustering scheme and destination aware context based routing protocol for VANET," Int. J. Intell. Networks, vol. 2, no. July, pp. 148–155, 2021, doi: 10.1016/j.ijin.2021.09.006.

[24] C. Wu, T. Yoshinaga, Y. Ji, and Y. Zhang, "Computational Intelligence Inspired Data Delivery for Vehicle-to-Roadside Communications," IEEE Trans. Veh. Technol., vol. 67, no. 12, pp. 12038–12048, 2018, doi: 10.1109/TVT.2018.2871606.

[25] I. Hussain and C. Bingcai, "Cluster Formation and Cluster Head Selection Approach for Vehicle Ad-Hoc Network (VANETs) using K-Means and Floyd-Warshall Technique," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 12, pp. 11–15, 2017, doi: 10.14569/ijacsa.2017.081202.

[26] A. B. Tambawal, R. M. Noor, R. Salleh, C. Chembe, and M. Oche, "Enhanced weight-based clustering algorithm to provide reliable delivery for VANET safety applications," PLoS One, vol. 14, no. 4, pp. 1–19, 2019, doi: 10.1371/journal.pone.0214664.

[27] J. Zheng, H. Tong, and Y. Wu, "A Cluster-Based Delay Tolerant Routing Algorithm for Vehicular Ad Hoc Networks," IEEE Veh. Technol. Conf., vol. 2017-June, pp. 1–5, 2017, doi: 10.1109/VTCSpring.2017.8108461.

[28] G. Husnain and S. Anwar, "An intelligent cluster optimization algorithm based on whale optimization algorithm for VANETs (WOACNET)," PLoS One, vol. 16, no. 4 April, pp. 1–22, 2021, doi: 10.1371/journal.pone.0250271.

[29] G. H. Alsuhli, A. Khattab, and Y. A. Fahmy, "Double-head clustering for resilient VANETs," Wirel. Commun. Mob. Comput., vol. 2019, 2019, doi: 10.1155/2019/2917238.

[30] M. S. Talib, A. Hassan, B. Hussin, Z. A. Abas, Z. S. Talib, and Z. S. Rasoul, "A novel stable clustering approach based on Gaussian Distribution and relative velocity in VANETs," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, pp. 216–220, 2018, doi: 10.14569/IJACSA.2018.090434.

[31] R. Monteiro, S. Sargento, W. Viriyasitavat, and O. K. Tonguz, "Improving VANET protocols via network science," IEEE Veh. Netw. Conf. VNC, pp. 17–24, 2012, doi: 10.1109/VNC.2012.6407428.

A Secure Unmanned Aerial Vehicle Service for Medical System to Improve Smart City Facilities

Birasalapati Doraswamy[1]*
Research Scholar, Department of Electronics and Communications Engineering
JNTUA College of Engineering Anantapur, Jawaharlal Nehru Technological University Anantapur (JNTUA)
Ananthapuramu, Andhra Pradesh-515002, India

K. Lokesh Krishna[2]
Professor, Department of Electronics and Communications Engineering
SV College of Engineering
Tirupati, Andhra Pradesh-517502, India

M.N. Giriprasad[3]
Professor, Department of Electronics and Communications Engineering
JNTUA College of Engineering Anantapur, Jawaharlal Nehru Technological University Anantapur (JNTUA)
Ananthapuramu, Andhra Pradesh-515002, India

***Abstract*—The use of drone technology and drones are currently widespread due to their increasing applications. However, there are some specific security-based challenges in the authentication process. In most drone-based applications, there are many authentication approaches, which are subject to handover delay issues with security complexities for an attack. To end these issues, the presented research has focused on developing a novel Optimized deep learning model known as Fruit Fly based UNet Drone Assisted Security (FFUDAS) to remove the malicious attacks. Moreover, the user requests are stored in the cloud, and the stored data are trained to the drones. Hereafter, the drones can deliver medicine to the requestor's location; in that, the malicious attacks were changes the location of drones. Once the attack is identified, then the attack removal process is done. Finally, the new path location to the requested user was identified with the help of fruit fly fitness; then the medicines are delivered to the requested user's location. Furthermore, the designed procedure is executed in an NS2 platform with required nodes. The robustness of the presented model was verified by evaluating the metrics like confidential data rate, execution time, handover delay, pack perception and data delivery rate, and energy consumption. Furthermore, to identify the effectiveness of the presented work, the presented model is compared with other existing schemes. The comparison results show that the presented model has higher throughput, less execution time and handover delay.**

***Keywords*—Drones; security; FFUDAS; malicious attack; fruit fly fitness; path identification; medicine delivery**

I. Introduction

In real-world applications, the technology used in a wide range was the Internet of Things (IoT) [1]. It consists of enormous objects that are interconnected through the environment [2]. The IoT objects are utilized to gather data from different sources and the collected data were exchanged over the internet [3]. This confirms that the objects within IoT make their own decisions without the need of humans [4]. Hence, the IoT's fundamental motivations are to integrate real-world physical and computerized systems for increasing the economic gains and to secure the information efficiently and accurately [5]. The drone was a type of flying IoT object or unmanned aerial vehicle (UAV); it was increasingly being deployed and developed across the globe [6]. Initially, these devices was used for military applications, but now these devices are adopted for different services in a wide range such as service delivery, traffic management, industrial monitoring, agriculture and healthcare [7].

The drone's IoT framework is shown in Fig. 1. These types of drones are currently used in IoT technology to play their part as Internet of drones (IoD), which comprises drones, remote users and ground station [8]. IoD has treated as a controlled architecture of layer network [9]. It was primarily developed to interconnect the UAVs access to support different navigation activities and control airspace [10]. In the growing number of IoT-enabled smart cities, there was widespread concern in using drones [11]. In smart cities, the most important issue was the authentication of drones during flight [12]. Thus, the drone with a secure network is important in each zones of the smart city [13].

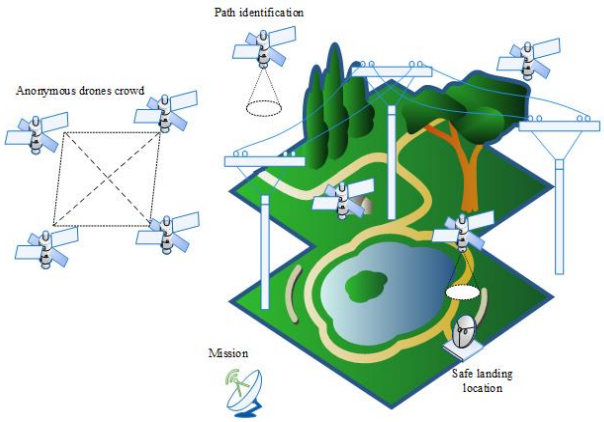


Fig. 1. Drones IoT Framework.

*Corresponding Author

Most importantly, security with low latency-based authentication mechanism is required for drone-assisted applications [14]. Moreover, preserving the service quality and eliminating the parameter effects may affect drones own mechanism [15].

In the IoD environment, the emerging application was drones in healthcare services. Using the healthcare drone services, the tasks like medicine delivery, medical equipment supply and collection of samples can be delivered to a particular area of patients [16]. Moreover, these services are also useful in tribe areas, the restriction imposed areas and rural areas [17]. However, the IoT healthcare service faces many privacy and security issues such as environmental-based attacks [18]. To secure the IoD environment, access control, key management and authentication are the primary services in security. Moreover, the utilization of blockchain technology makes the systems more robust against different attacks such as transparency, decentralization and immutability [19]. The information communicated through the IoT drone, the data were strictly confidential and private [20]. Several research works were done in the past such as authentication-based blockchain technology [21], 5G-based IoD environments blockchain challenges [22], drones system management and privacy [23], etc. are implemented in the past for secure sharing of things via UAV, but it gave poor outcomes due to inefficient algorithms and attack harmfulness. Hence the present work has aimed to develop a novel optimized deep learning methods in the UAV to enhance the monitoring function of malicious activities. By improving the monitoring function, the malicious attacks are identified and removed from the network environment. The main objective of this research is secure sharing of data.

The rest of the paper is described as follows: the recent literatures related to the drone for sharing things is described in Section 2, the system model and its problem statement is explained in Section 3, Section 4 explains the proposed methodology and its process, the result and discussion of the presented framework is described in Section 5 and Section 6 concludes the paper.

II. Related Work

Some recent literatures related to the drone for medicine sharing are described as follows:

The utilization of drone technology and the drone was widespread because of its rising applications such as safety surveillance, intelligent transportation, delivery, shipping, and military packages in IoT global landscape. However, the drone applications led to latency-based issues in real-time. To address this, Yazdinejad *et al.* [21] have presented a secure authentication-based model with less latency for drones which looks like leverage-based blockchain technology. Also, they implement an architecture zone in a drone network i.e., delegated drone stake proof. The results clearly demonstrated that the presented model has high throughput, less delay in end-to-end and less packet rate. Moreover, the energy consumption of drones is high and control of drone speed is difficult.

The 5G-based IoT-enabled Internet of Drones (IoD) environments blockchain applicability issues and in-depth challenges were presented by Bera et al. [22]. Moreover, IoD communication entities' data management's new blockchain secure framework was presented and analyzed. The result indicated that the presented method offers better functionality requirements and security. However, there were latency issues and threats are at high-level.

Due to the higher traffic demands of UAVs, it faces many challenges such as security, system management and privacy. To end this issue, Labib et al. [23] have presented a study about the UAV's current state-of-art and low-altitude traffic management in the airspace. It additionally explored the landscape technical standardization and highlighted the synergies among UAV operations standardization efforts and scientific research. The study result demonstrated that the IoT with drones has good privacy and security. Moreover, without guidelines, it does not identify the risk strategies.

Yahuza et al. [24] has assessed the recent trends in privacy and security issues, which affect the IoD-based network. Also, they investigated the privacy and security threat levels under various categories of the drone. The needed architecture for secured IoD networks and the comprehensive attacks taxonomy were highlighted. Moreover, the performance metrics and evaluation methods employed using techniques are also provided. However, many techniques face privacy-related issues and it was not rectified.

To tackle the Authenticated key management (AKM) issues in IoD environment, Tanveer et al. [25] have presented a robust AKM for IoD (RAMP-IoD), which utilizes lightweight cryptography. It also verifies the authenticity of users and it set a session-key among specific drones and users. The results indicated that the presented RAMP-IoD method had enhanced communication, computational overheads and high security. Furthermore, this protocol was not resource-efficient in IoD environmental security. The overall state-of-art comparison of existing literature is described in Table I.

The recent existing techniques did not resolve the security-based issues. Therefore, a novel nature inspired algorithm with network is designed in this research to resolve the security issues. Moreover, the key contribution of this research work is summarized as follows:

- Initially, the required number of nodes is designed in the NS2 environment.
- Consequently, a novel FFUDAS was designed to monitor the malicious activities in the present nodes.
- Hereafter, the malicious nodes are predicted and removed from the network environment.
- Thus, the drone-based IoT system was protected against the harmful activities; also the malicious nodes are in the way of data transfer, then data is handed over to other nodes by the fitness of fruit fly.
- Finally, the key metrics are calculated in terms of data delivery rate, confidentiality, handover delay, execution time, packet drop, energy consumption.

TABLE I. STATE-OF-ART COMPARISON

Sl.no	Authors	Techniques	Merits	Demerits
1	Yazdinejad *et al.* [21]	DDPOS	It can detect the attacks in an efficient manner with good accuracy	The energy consumption of drones is high and control of drone speed is difficult
2	Bera *et al.* [22]	BSD2C-IoD	It offers better functionality requirements and security	Latency issues and threats are in high-level
3	Labib *et al.* [23]	Study of UAV	The study reports that IoT enabled drone has precise results in security	This study does not provide future scopes and risk strategies
4	Yahuza *et al.* [24]	IoD security assessment	The performance of the drones security under variety of categories are determined	The privacy-based challenges are not resolved yet
5	Tanveer *et al.* [25]	RAMP-IoD	It improves the security, communication and performance	The process consumes more energy

III. SYSTEM MODEL AND PROBLEM DESCRIPTION

To improve the lifestyle of people and to reduce the human efforts and risks, drone application has become the most required sector in recent era. In drone communication securing the information is a much need task because the sensed data has remained in a wide range. So, it is vulnerable to get attacked by harmful malicious events.

Once the data is corrupted during the transmission, then the receiver or user can attain the wrong data, which is not useful for the specific users. Also, it might cause any wrong incident to that specific user. Moreover, in the medical field, security is the primary task; if the medical information from doctor to patient or from patient to doctor has got collapse, and then it tends to happen huge loss in the sense of money and health. The system model with problem in sharing is shown in Fig. 2.

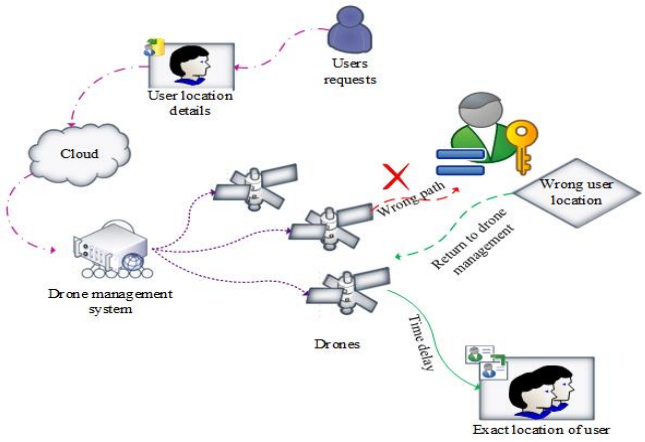


Fig. 2. System Model with Issues.

IV. PROPOSED METHODOLOGY

The drone-based medical delivery system is introduced in many rare situations to support the people from the tragedy. Hence, the medical assist drone contains user profile and location details. So to secure those data, the present research has aimed to design the security framework based on the monitoring and prediction model.

Moreover, the novel technique is named Fruit Fly based UNet Drone Assisted Security (FFUDAS) architecture was designed in the NS2 network. Subsequently, the malicious activities in between the drone were predicted and neglected from the drone environment. The proposed architecture is detailed in Fig. 3.

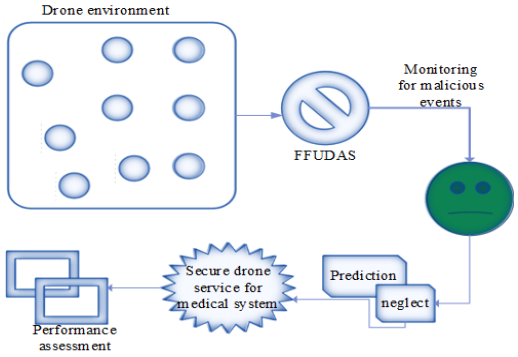


Fig. 3. Proposed Architecture.

A. Proposed FFUDAS Framework

The proposed FFUDAS model is a combination of Fruit fly optimization [26] and UNet based deep learning [27] approach. Initially, the cloud receives the user requirements and location, and is stored in the memory of UNet. Various blocks are designed to store the user details, but the current research work has focused on the supply of medicine in the emergency period. Moreover, from the requested data, the requested user is identified in the starting stage. Initially, the node selection is described in (1):

$$R(ny) = N^{*}\{1,2,3,.....,n\} \quad (1)$$

Where, $R(ny)$ represents the objective function, and N^{*} denotes the designed number of nodes in the network environment. After designing of nodes, the user needs and locations are collected, and then stored in the memory layer of the UNet. Hereafter, the information's are sent to the drone management or drone controller to supply the medicine to the requestor. The requested users' information gathering process is expressed in (2):

$$C_r^{*} = G.(U_r^{*}) \quad (2)$$

Where, C_r^{*} denotes the cloud receiver, G represents the gathering of user information, and U_r^{*} is the user needs. Moreover, the drone controller randomly initialized the location of the drone using a fruit fly-based location to send the

medicines. The fruit fly-based location identification is shown in (3):

$$D_{axis,i} = rand \times (D_{\max} - D_{\min}) + D_{\min} \quad (3)$$

Where, $D_{axis,i} = D_{axis,1}, D_{axis,2}, ..., D_{axis,n}$, $rand$ is the randomly generated drone in uniform and its range is [0,1], $D_{\max}$ and $D_{\min}$ are the distances of drone from the location of the requestor to drone management $i = 1,2,...,n$. Moreover, due to some attacks, the drone was moved to the wrong location. Hence, attacks in network is identified based on UNet and it is expressed in (4),

$$m(a) = m_x(a) + m_l . \exp\left(- \frac{(D_1(a) + D_2(a))^2}{2\beta^2} \right) \quad (4)$$

Where, $m(a)$ represents the presence of malicious attack, $m_x(a)$ denotes the moving location of drones through the attack, m_l denotes the drone pathway, $D_1(a)$ and $D_2(a)$ represents the distances between drones and attacks, and β is the range of drones. Based on the above expression, the attacks in the frames are identified. Moreover, after identification of attacks, it was removed by following (5),

$$A(r) = -\kappa\delta(m(a) - D_{axis,i}) \quad (5)$$

Where, κ represents the fitness of fruit fly, δ denotes the identified attack. Moreover, the proposed FFUDAS layer is shown in Fig. 4.

After the elimination of the attack, the new locations for the requestor are randomly generated through the search process, which is shown in (6),

$$D_{i,j} = D_{axis,i} + random \quad (6)$$

Where, $j = 1,2,.....N_{loc}$, N_{loc} represents the new location, and $random$ were range in the range of [-1,1]. Then the location of the user is calculated using (7),

$$U_{loc}^{*} = R(ny)_D \quad (7)$$

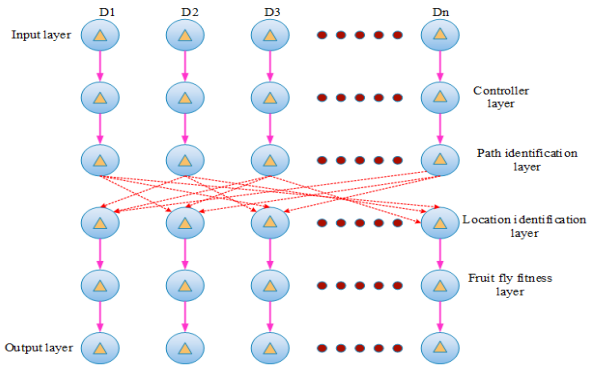


Fig. 4. Proposed FFUDAS Layer.

Here, U_{loc}^{*} is the location of user, and $R(ny)_D$ represents the searching process of locations. Moreover, by removing the attacks, the drones were successfully delivered the medicines to the requested user. Then the medicine-delivered drones are return to drone management. The pseudocode of the proposed FFUDAS framework is shown in Algorithm 1.

Algorithm 1: Proposed FFUDAS framework

```
Start
{
    Initialize: D1,D2,D3
    // here D1,D2,D3 are the used drones
      Information gathering process ()
       {
          user ⟶ requirements + location ⇒ S*m
        // here S*m denotes the storing and memory layer
    Analyze ⇒ requestor needs (subject)
     // analyzing the needs of requestor: requested things
       if (subject ⇒ medicine}
        {
          Medicine requested user
        }else (other things)
     Location identification process ()
     {
          int D → Daxis,i                    // using (3)
              // here D is the drone controller
     }
    Identification of attack ()
    {
        UAV → l*t + m(a)
        // here l*t represents the location of attack and m(a) represents the presence of attack
    }
       UAV working frame
       {
          int R, S;
            // here R is drones travel initiating point and S is the location of target
          Start (R) → End (S)
        start (R) =drone l*c; end (S) = requestor l*c
        // here l*c represents the location
     Drone ↔ requestor = initial ↔ ending
       // by defining the location of requestor, the initial and ending time of drone is fixed
       Return ⇒ starting point
       }
         Performance evaluation
}
    Stop
```

The travelling time of the drone from starting point to the target attaining point is determined using (8),

$$\eta^t = S_{ty} - E_{ty} \Rightarrow \min\_distance \quad (8)$$

Where, η^t represents the time taken to deliver the medicine to the requestor, S_{ty} denotes the drone starting time for medicine delivery, and E_{ty} represents the time at which the medicine was delivered in the correct location of the requestor. Furthermore, the drone speed depended on the weather condition, if the atmosphere has a high range of humidity, then the drone speed is low. Also, it takes more time to reach the target location. In addition, the flow chart of the proposed FFDUAS framework is shown in Fig. 5.

The presented FFUDAS model removes the malicious attacks from the nodes and creates a new path to deliver the medicine to the requested user. By removing the attacks, the paths were cleared and it delivered the medicine safely to the requestor location.

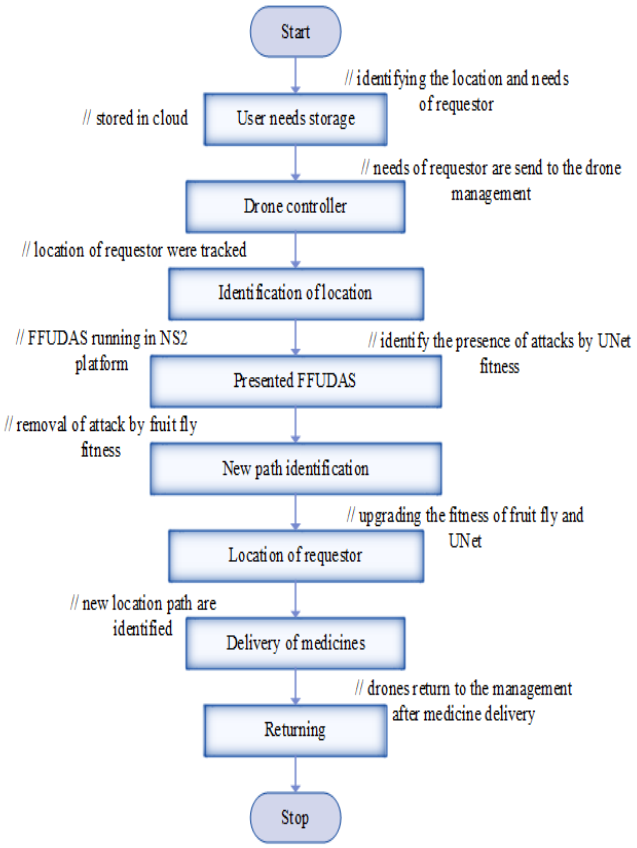


Fig. 5. Flow Chart of Proposed FFUDAS.

V. Result and Discussion

The presented research work is executed in the NS2 platform and running in the UBUNTU OS platform. Initially, the required number of nodes is designed with required labels; in these, some nodes are requested nodes (medicine), drone nodes, drone controller nodes and other normal nodes. In this process, initially, the requestor's needs and locations are stored in the memory of the cloud. Moreover, in the NS2 platform, the cloud memory is named as drone controller or drone management. The stored details are trained to the drones for medicine delivery to the target location of the requestor.

For numerical solutions, there are several optimizations, but the reason for selecting this particular fruit fly algorithm is to find the new path for the target location and to identify the location of attacks. The correlation of fruit fly and UNet removes the attack from paths and develops a new pathway. Generally, the fitness of the fruit fly algorithm is based on the location search. This reason has turned the interest to make use of fruit fly in this research.

A. Case Study

To validate the robustness of the presented model, 130 nodes are designed initially in the NS2 environment. According to this current research, the 130 nodes are considered as users and 6 nodes are requestors which are mentioned in sky blue color. The requestor needs are stored in the memory of the cloud that hub is colored as blue. The presented FFUDAS model node is represented in light brown color.

Moreover, to train the drones, drone management is required and it is represented in green color. Here, 3 nodes were totally used as a drone, which is mentioned in rose color. The node designed frame in NS2 is shown in Fig. 6.

Furthermore, after training the user location and requests to the drone, it has initiation the finding process of location, which is shown in Fig. 7. After the identification of paths, GPS helps to identify the location. The identified request or location frame is shown in Fig. 8.

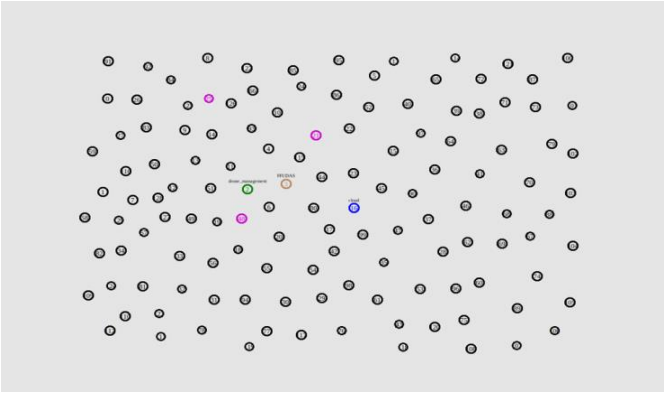


Fig. 6. Node Designed Frame.

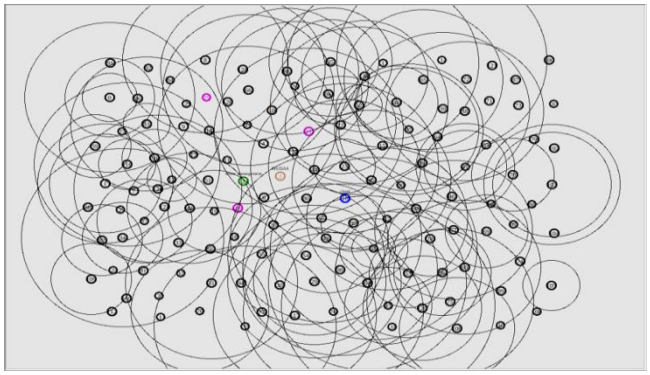


Fig. 7. Location Searching Frame.

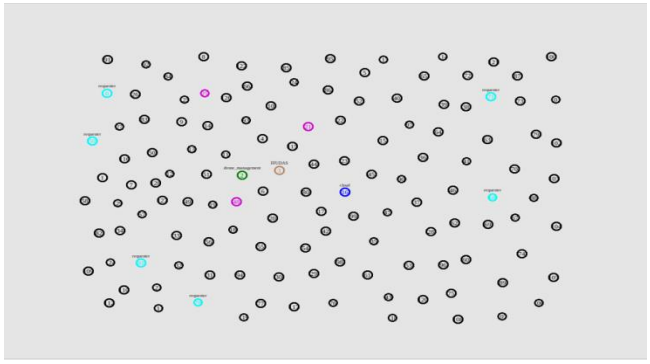


Fig. 8. Identified Requestor Location Frame.

After the identification of the requestor location, the drones were moved through the location. While the drone attempts to reach a zone, the malicious attack i.e., DoS (Denial of Service) like AUTH attack was happened. Moreover, the presented DoS attack was mentioned in red color and it is removed by the presented FFUDAS model. Moreover, the designed attacks and removed attacks in the designed frame are shown in Fig. 9 and 10, respectively.

Finally, the initiated drones have reached the destination and delivered the medicines after the removal of attacks which is represented in Fig. 11. When the things delivered, then the drone returns to the location of drone management, where it has been started.

The green color indicates the medicine delivered to the particular requestor. After delivery of medicines, the drones return to the drone management is represented as pink color.

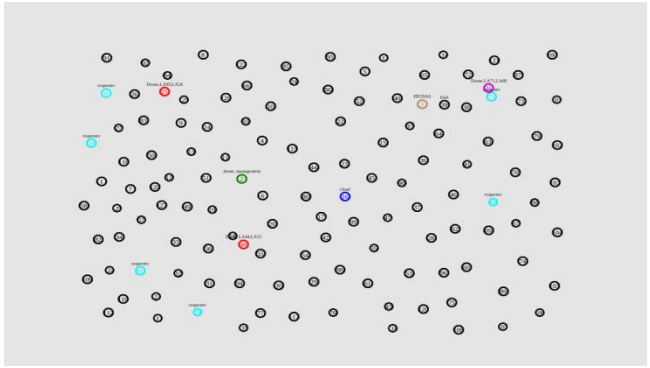


Fig. 9. Attacks in Designed Frame.

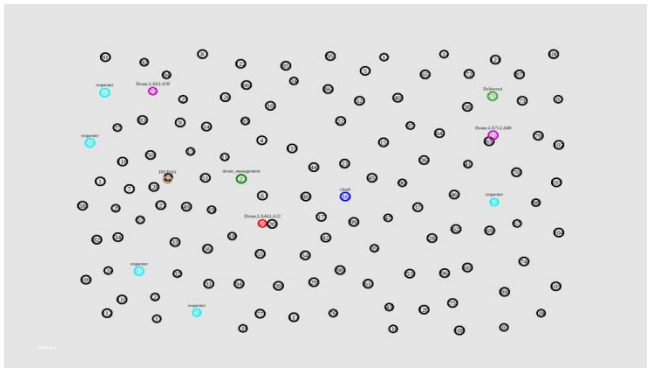


Fig. 10. Removed Attacks in Frame.

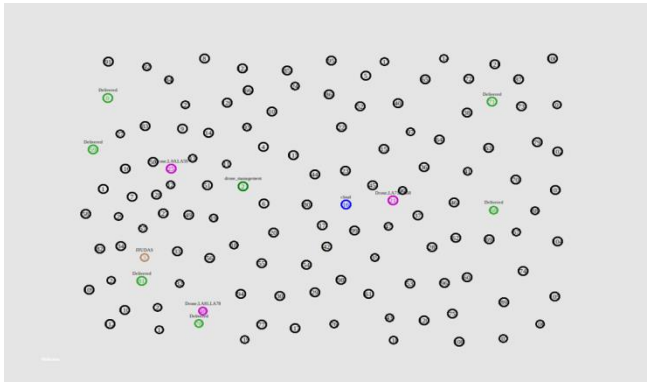


Fig. 11. Request Delivered.

B. Performance Evaluation

To validate the presented FFUDAS models proficient score, the function validation is a crucial task. Moreover, the robustness and working of the designed model were analyzed by evaluating the key metrics with different data counts. Hence, the metrics like data delivery rate (DDR), confidential rate, handover delay, execution time, packet reception rate (PRR), and energy consumption were validated for the different data counts.

1) Data delivery rate (DDR) and packet perception rate (PPR): Data delivery rate is defined as the ratio of a difference between the number of data sent and the number of data bounces to the number of data sent. It is calculated using (9),

$$DDR = \frac{N_{ds} - N_{db}}{N_{ds}} \quad (9)$$

Where, N_{ds} represents the total number of sent data and N_{db} denotes the total number of data bounces. The obtained DDR are shown in Fig. 12 and the results obtained are shown in Table II.

The packet perception rate is defined as the ratio of the number of packets delivered in the target location to the number of packets sent to the target location. It is expressed in (10):

$$PPR = \frac{\sum D_{p.t}}{\sum S_{p.t}} \quad (10)$$

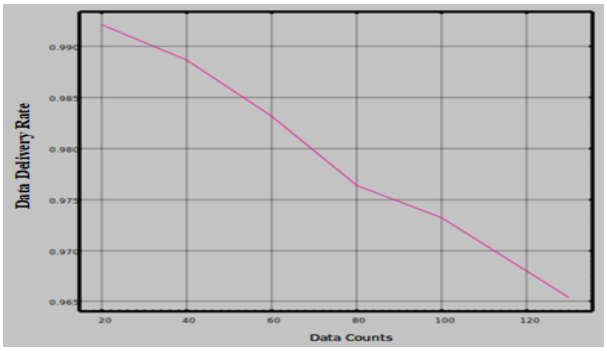


Fig. 12. Data Delivery Rate.

TABLE II. OBTAINED DATA DELIVERY AND PACKET PERCEPTION RATE

Sl.no	Data count	DDR	PPR
1	20	0.9921	0.98
2	40	0.9887	0.975
3	60	0.9832	0.951
4	80	0.9764	0.932
5	100	0.9732	0.925
6	130	0.9654	0.918

Where, $\sum D_{p.t}$ represents the total number of packets delivered to the requestor and $\sum S_{p.t}$ represents the total number of packets sent to the requestor. Moreover, the obtained PPR at different data counts is shown in Fig. 13.

The results indicated that the presented model has higher DDR and PPR at 130 nodes. The proposed framework has the finest result in both data delivery and packet perception rate. The DDR and PPR has attained mean as 0.98, and 0.95, respectively, which are effective for successive data sharing.

2) Execution time and Handover delay: The length of the time needed to perform a complete process is known to be execution time. It is also known as computation time or running time. Moreover, it is proportional to the rule applications. Its unit is meter second and the time obtained to complete the process in different data counts is shown in Fig. 14.

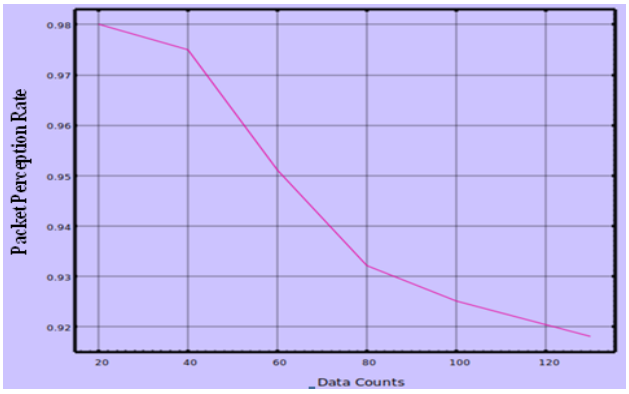


Fig. 13. Obtained PPR at different Data Counts.

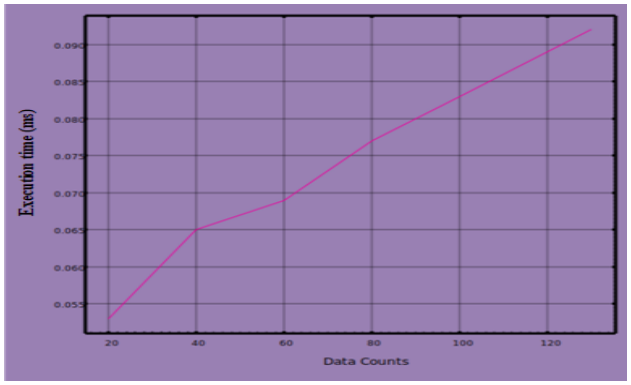


Fig. 14. Execution Time in different Data Counts.

Due to attacks, the delivery of the material was delayed to the requestor from the setting time. However, the presented FFUDAS framework has less handover delay due to removal of attacks in the network. Furthermore, the obtained result of execution time and handover delay is shown in Table III.

TABLE III. RESULT OF EXECUTION TIME AND HANDOVER DELAY

Sl.no	Data count	Execution time (ms)	Handover delay (ms)
1	20	0.053	0.2
2	40	0.065	0.25
3	60	0.069	0.39
4	80	0.077	0.4
5	100	0.083	0.5
6	130	0.092	0.8

Moreover, handover delay is the time, which taken for redirecting the ongoing location, when the node changes its point from one location to another. The obtained handover delay is shown in Fig. 15 and its unit is meter second (ms).

3) Confidential rate and energy consumption: In contrast, when the drone is close to the requested user, the drone usually slows down and it increases the transmission power for the confidential data rate. The confidential data rate is high at 20 data counts and less at 130 data counts. Moreover, the test results of confidential rate and energy consumption at different data counts is shown in Table IV.

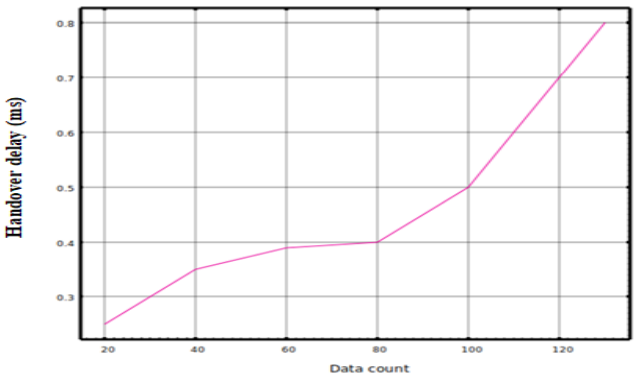


Fig. 15. Handover Delay at different Data Counts.

TABLE IV. RESULT OF ENERGY CONSUMPTION AND CONFIDENTIAL RATE

Sl.no	Data count	Confidential data rate (Mbps)	Energy consumption (Joule) $\times 10^4$
1	20	130	2.10
2	40	125	2.12
3	60	120	2.16
4	80	110	2.18
5	100	100	2.22
6	130	95	2.25

Furthermore, the accumulated test result of confidential data rate and the energy consumption is shown in Fig. 16 and 17 respectively. The result indicated that the presented model has less energy consumption in all the nodes and the confidential rate were decreased gradually due to the elimination of attack nodes.

In UAVs, the important part is energy consumption; the drone with less energy consumption only reaches the target location without any delay. The energy of drones receives through wireless charging, which is utilized to process the user's tasks in the drone coverage area. The unit of energy consumption is joule.

Energy consumption is evaluated for determining how much energy is being consumed by the FFUDAS framework. The result demonstrated that the presented framework consumes less energy over a large distance. Normally, due to changes in weather conditions, the drone consumes higher energy.

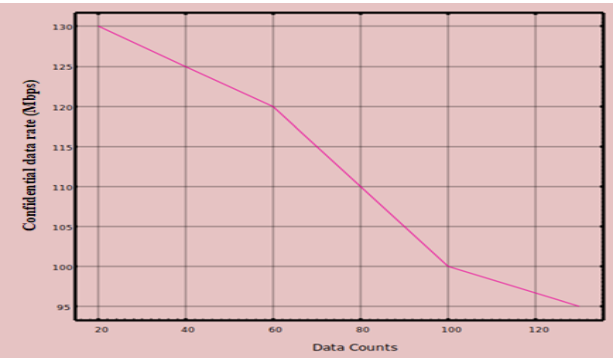


Fig. 16. Confidential Data Rate.

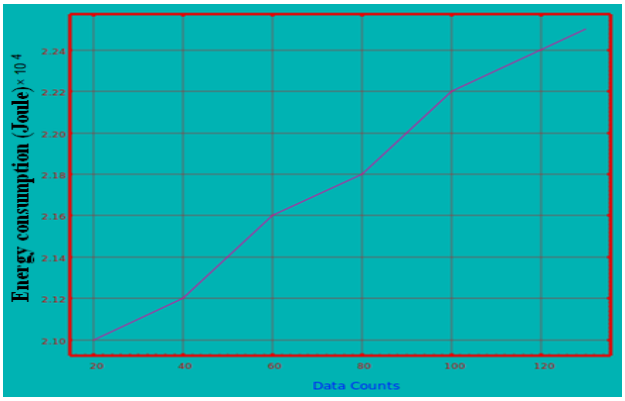


Fig. 17. Energy Consumption.

C. Comparative Analysis

The performance of any application can be valued by validating the chief metrics like throughput, execution time, and handover delay. To identify the effectiveness of the presented research work, the presented model was compared with other existing models like DDPOS [21], BSD2C-IoD [22], BACS-IoD [28], and SDN-MIH [29]. The comparison result is shown in Table V.

The comparison result indicated that the presented model has higher throughput than other models. The presented FFUDAS model has attained the throughput of 0.96 Mbps, DDPOS has attained the throughput as 0.000275 Mbps, and SDN-MIH-UAV has attained 0.167 Mbps as throughput. The comparison of throughput and handover delay is shown in Fig. 18.

From the comparison result, the attained handover delay of the proposed FFUDAS model was 0.423ms, DDPOS was 0.425ms, and SDN-MIH-UAV was 3.02ms. The result indicated that the presented model has less handover delay than other existing works. Moreover, the comparison of execution time with other existing models is shown in Fig. 19.

Moreover, the yielded execution time of the presented FFUDAS model was 0.073ms, DDPOS was 5.57ms, BSD2C-IoD was 0.97 and BACS-IoD was 1.33ms. The comparison result demonstrated that the presented model has less execution time than other models.

TABLE V. COMPARISON OF METRICS WITH EXISTING WORKS

Sl.no	Techniques	Throughput (Mbps)	Execution time (ms)	Handover delay (ms)
1	DDPOS [21]	0.000275	5.57	0.425
2	BSD2C-IoD [22]	-	0.97	-
3	BACS-IoD [28]	-	1.33	-
4	SDN-MIH-UAV [29]	0.167	-	3.02
5	Proposed (FFUDAS)	0.96	0.073	0.423

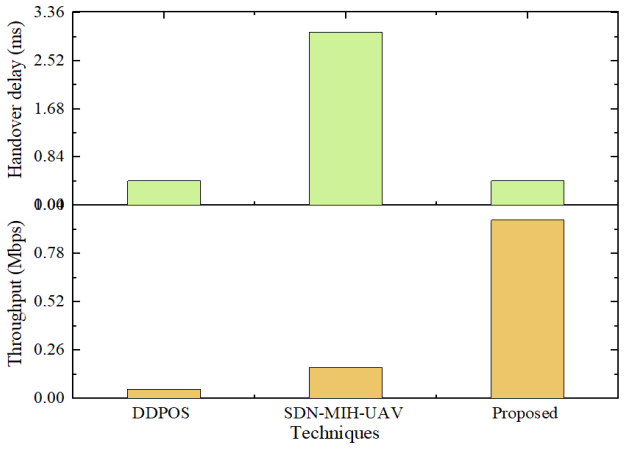


Fig. 18. Comparison of Throughput and Handover Delay.

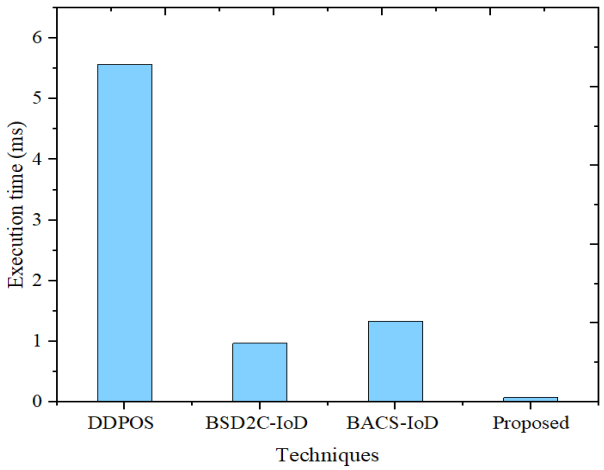


Fig. 19. Comparison of Execution Time.

D. Discussion

The outcome of the designed model has gained the finest results in all key metrics, which shows the effectiveness of the model. Also, the present research method's plan will help to enlarge the smart city applications. Moreover, the presented model has less execution time of 0.073ms. So, considering that the proposed FFUDAS technique has gained the best outstanding results within a short duration. The proposed FFUDAS overall performances mean is shown in Table VI.

TABLE VI. OVERALL PERFORMANCE OF FFUDAS

Performance of FFUDAS	
Parameters	**Obtained score**
DDR	0.98
PPR	0.95
Execution time (ms)	0.073
Handover delay (ms)	0.423
Confidential data rate (MbPS)	113
Energy consumption x 10^4 (Joule)	2.17

Hence, the presented model is applicable to supply medicines through drones and adoptable for monitoring and removing attacks. The novel FFUDAS takes average 0.98 as DDR that is high; the techniques with high data delivery rate are applicable for smart city applications.

VI. CONCLUSION

To ensure secure communication and less delay between UAVs in a smart city, the process of authentication must be established properly between the requestors in each zone. In some cases, the security of drones is a complex problem due to attacks. Moreover, the presented research has developed an FFUDAS model to remove and identify the attacks. The fitness of fruit fly is upgraded in the location identification layer to gain the finest results. Finally, the robustness of the presented model was evaluated by measuring the metrics like data delivery and packet perception rate, confidential data rate, execution time, energy consumption and handover delay. In all metrics, the presented FFUDAS model has yielded better results by attaining 0.96 Mbps throughput, 0.073 ms execution time and 0.423 ms hand over delay. By comparing with other models, the presented model has attained the finest outcome.

REFERENCES

[1] A. Khanna, and S. Kaur, "Internet of Things (IoT), applications and challenges: A comprehensive review," Wirel. Pers. Commun. vol. 114, pp. 1687-1762, 2020.

[2] A. Aslam, U. Mehmood, M. H. Arshad, A. Ishfaq, J. Zaheer, A. Ul Haq Khan, and M. Sufyan, "Dye-sensitized solar cells (DSSCs) as a potential photovoltaic technology for the self-powered internet of things (IoTs) applications," Sol. Energy, vol. 207, pp. 874-892, 2020.

[3] R. Yugha, and S. Chithra, "A survey on technologies and security protocols: Reference for future generation IoT," J. Netw. Comput. Appl. pp. 102763, 2020.

[4] M. Javaid, A. Haleem, R. Vaishya, S. Bahl, R. Suman, and A. Vaish, "Industry 4.0 technologies and their applications in fighting COVID-19 pandemic," Diabetes Metab. Syndr.: Clin. Res. Rev. vol. 14, no. 4, pp. 419-422, 2020.

[5] M. M. Nair, S. Kumari, and A. K. Tyagi, "Internet of Things, Cyber Physical System, and Data Analytics: Open Questions, Future Perspectives, and Research Areas," Proceedings of the Second International Conference on Information Management and Machine Intelligence, Singapore: Springer, 2021.

[6] F. Al-Turjman, M. Abujubbeh, A. Malekloo, and L. Mostarda, "UAVs assessment in software-defined IoT networks: An overview," Comput. Commun. vol. 150, pp. 519-536, 2020.

[7] T. Alladi, V. Chamola, N. Sahu, and M. Guizani, "Applications of blockchain in unmanned aerial vehicles: A review," Veh. Commun. vol. 23, pp. 100249, 2020.

[8] Z. Ali, S. A. Chaudhry, M. S. Ramzan, and F. Al-Turjman, "Securing smart city surveillance: A lightweight authentication mechanism for unmanned vehicles," IEEE Access, vol. 8, pp. 43711-43724, 2020.

[9] P. Boccadoro, D. Striccoli, and L. A. Grieco, "An extensive survey on the Internet of Drones," Ad Hoc Netw. vol. 122, pp. 102600, 2021.

[10] C. Xu, X. Liao, J. Tan, H. Ye, and H. Lu, "Recent research progress of unmanned aerial vehicle regulation policies and technologies in urban low altitude," IEEE Access, vol. 8, pp. 74175-74194, 2020.

[11] Z. Qadir, F. Ullah, H. S. Munawar, and F. Al-Turjman, "Addressing disasters in smart cities through UAVs path planning and 5G communications: A systematic review," Comput. Commun. 2021.

[12] S. Garg, G. S. Aujla, A. Erbad, J. J. P. C. Rodrigues, M. Chen, and X. Wang, "Guest Editorial: Blockchain Envisioned Drones: Realizing 5G-Enabled Flying Automation," IEEE Netw. vol. 35, no. 1, pp. 16-19, 2021.

[13] H. Teng, M. Dong, Y. Liu, W. Tian, and X. Liu, "A low-cost physical location discovery scheme for large-scale Internet of things in smart city through joint use of vehicles and UAVs," Future Gener. Comput. Syst. vol. 118, pp. 310-326, 2021.'

[14] M. Aloqaily, O. Bouachir, A. Boukerche, and I. Al Ridhawi, "Design guidelines for blockchain-assisted 5G-UAV networks," IEEE Netw. vol. 35, no. 1, pp. 64-71, 2021.

[15] A. S. Khan, G. Chen, Y. Rahulamathavan, G. Zheng, B. Assadhan, and S. Lambotharan, "Trusted UAV network coverage using blockchain, machine learning, and auction mechanisms," IEEE Access, vol. 8, pp. 118219-118234, 2020.

[16] L. Abualigah, A. Diabat, P. Sumari, and A. H. Gandomi, "Applications, Deployments, and Integration of Internet of Drones (IoD): A Review," IEEE Sens. J. 2021.

[17] D. S. Prashanth, J. Swamy, and S. S. Rao, "Internet of Things and Web Services for Handling Pandemic Challenges," Sustainability Measures for COVID-19 Pandemic, Singapore: Springer, 2021, pp. 1-19.

[18] S. Singh, P. K. Sharma, B. Yoon, M. Shojafar, and G. H. Cho, "Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city," Sustain. Cities Soc., vol. 63, pp. 102364, 2020.

[19] B. Bera, A. K. Das, and A. K. Sutrala, "Private blockchain-based access control mechanism for unauthorized UAV detection and mitigation in Internet of Drones environment," Comput. Commun. vol. 166, pp. 91-109, 2021.

[20] J. Chen, W. Wang, Y. Zhou, S. H. Ahmed, and W. Wei, "Exploiting 5G and blockchain for medical applications of drones," IEEE Netw. vol. 35, no. 1, pp. 30-36, 2021.

[21] A. Yazdinejad, R. M. Parizi, A. Dehghantanha, H. Karimipour, G. Srivastava, and M. Aledhari, "Enabling drones in the internet of things with decentralized blockchain-based security," IEEE Internet Things J. vol. 8, no. 8, pp. 6406-6415, 2020.

[22] B. Bera, S. Saha, A. K. Das, N. Kumar, P. Lorenz, and M. Alazab, "Blockchain-envisioned secure data delivery and collection scheme for 5g-based iot-enabled internet of drones environment," IEEE Trans. Veh. Technol. vol. 69, no. 8, pp. 9097-9111, 2020.

[23] N. S. Labib, M. R. Brust, G. Danoy, and P. Bouvry, "The Rise of Drones in Internet of Things: A Survey on the Evolution, Prospects and Challenges of Unmanned Aerial Vehicles," IEEE Access, vol. 9, pp. 115466-115487, 2021.

[24] M. Yahuza, M. Y. I. Idris, I. B. Ahmedy, A. W. A. Wahab, T. Nandy, N. M. Noor, and A. Bala, "Internet of Drones Security and Privacy Issues: Taxonomy and Open Challenges," IEEE Access, vol. 9, pp. 57243-57270, 2021.

[25] M. Tanveer, N. Kumar, and M. M. Hassan, "RAMP-IoD: A Robust Authenticated Key Management Protocol for the Internet of Drones," IEEE Internet Things J. 2021.

[26] Y. Fan, P. Wang, A. A. Heidari, M. Wang, X. Zhao, H. Chen, and C. Li, "Rationalized fruit fly optimization with sine cosine algorithm: a comprehensive analysis," Expert Syst. Appl. vol. 157, pp. 113486, 2020.

[27] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020.

[28] B. Bera, D. Chattaraj, and A. K. Das, "Designing secure blockchain-based access control scheme in IoT-enabled Internet of Drones deployment," Comput. Commun. vol. 153, pp. 229-249, 2020.

[29] S. Goudarzi, M. H. Anisi, D. Ciuonzo, S. A. Soleymani, and A. Pescapé, "Employing Unmanned Aerial Vehicles for Improving Handoff using Cooperative Game Theory," IEEE Trans. Aerosp. Electron. Syst. vol. 57, no. 2, pp. 776-794, 2020.

A Channeled Multilayer Perceptron as Multi-Modal Approach for Two Time-Frames Algo-Trading Strategy

Noussair Fikri, Khalid Moussaid, Mohamed Rida, Amina El Omri, Noureddine Abghour
LIMSAD Labs, Faculty of Sciences, Hassan II University of Casablanca
Km 8 Route d'El Jadida,Casablanca, Morocco

***Abstract*—FOREX (Foreign Exchanges) is a 24H open market with an enormous daily volume. Most of the used Trading strategies, used individually, are not providing accurate signals. In this paper, we are proposing an automated trading strategy that fits random market behaviors. It is based on neural networks applying triple exponential weighted moving average (EMA) as a trend indicator, Bollinger bands as a volatility indicator, and stochastic RSI as a momentum reversal indicator to prevent false indications in a short time frame. This approach is based on trend, volatility, and momentum reversal patterns combined with a market adaptive and a distributed multi-layer perceptron (MLP). It is called channeled multi-layer perceptron (CMLP) that is a neural network using channels and routines trained by previous profit/loss earned by triple EMA crossover, Bollinger Bands, and Stochastic RSI signals. Instead of using classic computations and Back-propagation for adjusting MLP parameters, we established a channeled multi-layer perceptron inspired by a multi-modal learning approach where each group of modalities (Channel) has its K_c That stands for a dynamic channel coefficient to produce a multi-processed feed-forward neural network that prevents uncertain trading signals depending on trend-volatility-momentum random patterns. CMLP has been compared to Multi-Modal GARCH-ARIMA and has proven its efficiency in unstable markets.**

Keywords*—FOREX; neural networks; EMA; Bollinger band; stochastic RSI; momentum reversal; MLP; back-propagation; feed-forward

I. Introduction

Financial markets are where securities trades occur, including the forex market, stock market, bond market, and derivatives market. Financial markets are the primary source of liquidity for businesses. Our paper focused on the most common and liquid financial market, which is the foreign exchange market. F.X., Forex, or Foreign Exchange trades one currency for another, for example, GB Pound vs. U.S. Dollar, known as GBP/USD instrument[1]–[3]. Forex has no physical location; it's an electronic market, technically an extensive network of financial institutions (banks and brokers) and individual traders that operate through brokers or banks. There is a significant amount of data generated by market prices movements. From the big data view, it represents a rich source of features for predictions[3]. Algorithmic trading implements manually used strategies to predict when prices go down and prices go up, then place buy or sell trades with significant consideration of taking profit and stopping loss parameters for the best portfolio management. Machine learning is the best way to build a robust algorithmic trading platform by retrieving historical market data, selecting efficient features then training machines for future market movement prediction[1], [4]–[6]. To predict market prices, traders use several technics to have an accurate expectation on future trending. The most used technics are based on detecting trading signals from trends [5], [7]–[9], Volatility, or Momentum behavior[10]–[14]. For trend, the well-known technic is moving average; it has three famous alternatives; the first is Simple Moving Average[1], [15], [16], it uses the last N values average. An enhanced version of moving average is exponential weighted moving average[17], [18]; instead of assigning the same weight to all values, EMA gives more weight to last values. This technique has evolved, and traders now use a triple EMA with different selective periods to predict market prices accurately. Especially when there is a crossover between the 3 EMA lines, it's interpreted as a signal to place a trade that is not consistently profitable, and it's called a false signal. False or True profitable signals follow a different pattern that distinguishes their behavior. For volatility, the well-known method is the Bollinger Band which uses the Simple Moving Average as anchor line and the standard deviation as an indicator of aggressive markets movement[12], [19], [20]. Then finally momentum that has a famous method, the Stochastic Relative Strength Index that shows the market position reversal[21]–[23].

To discover the winning patterns based on cited indicators, we need to parse historical market data and train a neural network as the chosen algorithm of the machine learning branch. There are a lot of neural network types. The convolutional neural network, which belongs to the deep learning class, is used for visual imagery[24]–[26], and the recurrent neural network, which performs in natural language processing[18], [27]–[29]. In our case, we are dealing with a dataset of time series data (Financial market prices) processing and produce a binary decision: Place a profitable trade or Place none profitable trade[30], [31]. We choose a MultiLayer perceptron as a feed-forward neural network by using a back-propagation algorithm[24], [32]. On our approach, the features are calculated previously from historical data. We combined 3 EMAs and their growth rates as trend features, Bollinger Bands as volatility features, and Stochastic RSI as momentum features to enhance the accuracy of predicted decisions from current patterns and follow an adaptive model based on

different criteria (Trend- Momentum - Volatility). Let's go back to the existing state of the art. Several methods offer almost the same deep learning architecture but ignore the leading indicators of the instability of a specific market. Among these methods, we will cite the way that predicts price changes using LSTM [33]. Then a compound method consists of producing features based on GARCH to frame the data that describe volatility and their link to price variation and then inject them as input to an LSTM like N.N. to predict the subsequent variations. We will also cite a mixed method to prepare trend and seasonality data using the ARIMA approach and then inject data on an LSTM to predict the next trend. And finally, couple the last two methods quoted on a Multi-modal architecture, where the first modal is the LSTM-GARCH, and the 2nd modal is the LSTM-ARIMA. That will allow us to create the perfect hybrid model to benchmark our Channeled Multilayer Perceptron approach.

II. Related Work

A. Basic Trading Strategies

Day trading is an activity that involves entering winnings of less than ten pips per transaction on windows with 1, 5, and 15 minutes; this technique is called scalping. The simple moving average is the most popular among the statistical methods used to predict future prices, and it's an essential technique to describe a specific market trend [1], [15]. The first method we will study is the Simple Moving Average (SMA) which is the average of the previous N last prices, as shown on the following statement:

$$SMA = \frac{\sum_{i=0}^{N} Price_i}{N}$$

Then we have an Exponential moving average (EMA) that works the same as a simple moving average, except it places greater weight on the more recent closing prices[13], [15], [17].

$$M = \frac{2}{N+1}$$

$$EMA = M \times (ClosingPrice - EMA(Previous\ Day)) + EMA(Previous\ Day)$$

The difference between the simple and exponential moving average is that the EMA is smoother than the SMA; it is safer to use the EMA as a trend indicator. However, there is a more accurate technique to predict the trend of a specific financial instrument. It is called EMA crossover. It can be used with 2 or 3 different periods and detect their curve crossing. In the case of 2 respective curves with two periods, if the shortest EMA period crosses the more extended EMA period from bottom to top, it is considered an uptrend signal. If it passes from top to bottom, it is regarded as a downtrend. In the case of 3 periods, if the shortest EMA period crosses the long EMA period from bottom to top and both are above the longer EMA period; this is considered a more accurate up-trend signal. And if it goes from top to bottom and both are below the longer EMA period, this is considered a more accurate down-trend trend signal.

We also have a Moving Average Divergence and Convergence (MACD) trading strategy that uses 3 AMS with fixed windows of 26 EMA periods, 12 EMA periods, and 9 EMA periods overlapping [21] [22]. It reads positive and negative movements based on a zero line. The buy signal is triggered when the MACD passes over the zero line from the bottom, considered an uptrend, and sells where it crosses the signal line below, considered a downtrend. The take profit is triggered when the MACD falls below the signal line if a purchase order is placed. And if a sell order is placed, the profit-taking signal is triggered when the MACD rises above the signal line. The momentum of time series assumes that a market that behaves poorly with losses or profits over specific periods will continue at the same rate. It is based on the average log of N periods and its measure of performance above or below a red line on which the position is defined as positive or negative.

For the measurement of volatility, the chosen method is the Bollinger band which uses the simple moving average as an anchor line and the standard deviation as an indicator of aggressive market movement. It has been developed and protected by the copyright of John Bollinger, a trading expert, to provide meaningful insights into the specific volatility of the market [12], [19]–[21]. Thus, the first step in the calculation of the Bollinger band is to obtain the typical prices of the chosen period, then the standard deviation over the same period, and finally to calculate them with the simple moving average of the upper and lower bands, as follows:

$$BBUpper = SMA(TP, n) + \left(m \times \sigma(TP, n)\right)$$

$$BBLower = SMA(TP, n) + \left(m \times \sigma(TP, n)\right)$$

With:

- BBUpper: Upper Bollinger Band
- BBLower: Lower Bollinger Band
- SMA: Simple Moving average
- TP (typical price): (High+Low+Close) ÷3
- N: Number of days in smoothing period (typically 20)
- m: Number of standard deviations (typically 2)
- σ[T.P.,n]Standard Deviation over last n periods of T.P.

Fig. 1 shows the produced Bollinger lower and upper bands using 20 as a number of smoothing days and two as standard deviation:

Fig. 1. Plot of Bollinger Band using as 20 Smoothing Periods and 2-Period Standard Deviation.

And finally, The Stochastic RSI (Stochastic Relative Strength Index) is an application of a stochastic oscillator on a set of RSI[22], [23]. It varies between 0 and 1 or 0 and 100 percent to indicate an overbought or oversold of a specific market, in other words, a momentum position reversal. It is calculated on N-period by using two steps as follows:

$$RSI = 100 - \left[\frac{100}{1 + \frac{Previous\ average\ gain * N + Current\ gain}{Average\ average\ loss * N + Current\ lossPrevious}}\right]$$

Then:

$$StochRSI = \frac{RSI - MIN(RSI, N)}{MAX(RSI, N) - MIN(RSI, N)}$$

Fig. 2 shows the produced 80-Overbought and 20-Oversold lines of a 14-period stochastic RSI:

B. An Overview

1) Existing approach for prediction: There are several machine learning techniques, ranked according to their objectives, such as fault detection for unusual data points using One-class SVM [16], [33], [34], which can be a good solution if there are more than 100 features, with aggressive boundaries. We can also use the detection of anomalies by PCA in case we need fast training [13], [35], [36], or Clustering to discover the structure using the algorithm K-means if we want to label data that belong to a specific category or group [14]. But this cannot be used in our study because any market behavior is not necessarily an anomaly; the appropriate indicator must consider any aberrant data. Otherwise, for the prediction of continuous values, the best approach is linear regression if we follow a linear model [4], [14], [37], otherwise, for a two-class or multiclass binomial prediction, the classic approach is logistic regression, but we prefer to use a neural network.

Fig. 2. Plot of 14-Period Stochastic-RSI on Short-Time Frame.

For the classification to two classes and multiclass [38], [39], we can mention the decision tree, which has the ability of interpretation and categorization by following a binary or multiclass configuration [40], [41]. It can even use random forests and gradient-boosted tree algorithms as sets of trees to optimize classification and regression processes [42], [43]. There is also the Support Vector Machine or Networks (SVM) [16], [34], [44], which is a combination of supervised learning models and trained learning algorithms. It depends on partitions of spaces of high or infinite dimensions defined by parallel lines called hyperplanes, or maybe a set of hyperplanes with an operating margin quantified by distance to learning data points; the smaller the margin, the greater the generalization error of the classifier. Other methods, such as the One-vs-Rest classifier and naive Bayes, do not correspond to our case study.

2) Convolution neural network: In complex images, for example, face recognition, pictures have pixel dependencies where CNN is more accurate in prediction [24], [26]. Therefore, the prominent role of the convolutional network is reducing the complexity of images input features without losing accuracy. As shown in Fig. 3, a convolutional neural network is composed of, Convolution Layer, Pooling Layer, Flatten Layer, Fully connected layer, and SoftMax Layer [24], [25], [26], [30]:

3) Recurrent neural network: RNN is a feed-forward neural network that remembers the past. It takes what is learned from previous training iterations in its memory and uses it for the next predictions. RNN accepts one or more input vectors then produces one or more output vectors. It's not limited to associated weights but also hidden vectors which are considered as its memory. For each X input with an associated weight W_x there is a hidden node H with an associated weight W_y and an output Y with associated weight Y. A recursive process is applied on hidden states, accurately on H weights and remembering H and W_h values. The exact process is used on deep RNNs, weights are recursively updated in each iteration, and then more iterations are running, we go deep in learning, the reason we call it deep RNNs[18], [27]–[29].

4) Basic multilayer perceptron: MLP applies activation function where the primary is introducing the non-linearity to neural network outputs. Adding bias to this function provides additional trainable weight to a fixed value (Constant) that shifts output curves depending on the inputs group[30]. These are three known activation functions defined by the following separated expressions:

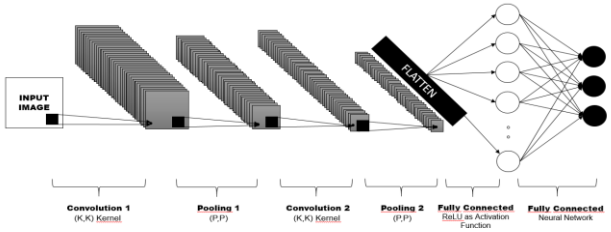


Fig. 3. Full Representation of Convolutional Neural Network.

$$Z = \sum_{i=0}^{n} W_i x_i + B$$

The Sigmoid function that computes the real value to range it between 0 and 1 as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The tanH function that computes the real value to range it between -1 and 1 as follows:

$$\tanh(z) = 2\sigma(2z) - 1$$

The ReLU, which means Rectified Linear Unit, computes real-values to replace negative values by 0 as follows:

$$F(z) = \max(0, z)$$

Feed-forward NN is the simplest form of ANNs [6], [23], [28]. It's composed of multiple neurons or nodes organized as layers with connections or edges between them containing weights of each node. After the first iteration comes the back-propagation algorithm to adjust the random weights associated with N.N. training ignition. By knowing the targeted value and the outputted value of the activation function, we get the difference between them as an error value to use back-propagation to calculate the gradients and adjust weights as optimal as possible by using a derivative function [6], [18], [26], [32], until error is reduced as shown on the following Fig. 4:

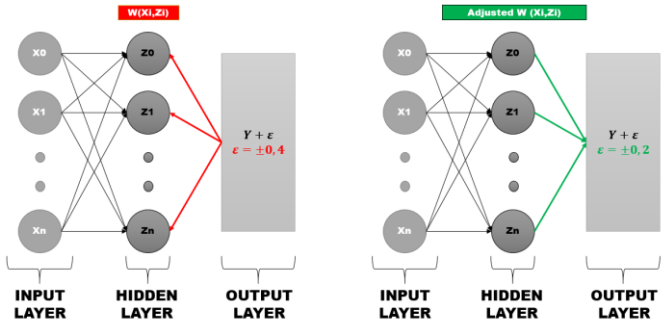


Fig. 4. A Preview of the Back-Propagation Process.

C. Available also-Trading Approaches

Several proposed approaches combine machine learning and forex market prices prediction, especially neural networks, to predict a trend of a specific instrument (e.g., GBP/USD). Most of them choose multi-layer perceptron, which seems to have the necessary ability and accuracy to recognize an effective trending pattern. Zhai, Hsu, and Halgamuge (2007) [8] choose a combination of market-related news and technical indicators delivered as features to a Support Vector Machine (SVM) based prediction model. Leslie Tiong Ching Ow [4] found a technic of trend prediction based on considering linear regression line (LRL) of the prices of timeline and trend line, in other words, the distance between the two lines, as a pattern of a possible up-trend or down-trend. Then use these patterns to train a Multilayer Perceptron for predicting future market trend movement. Gene I. Sher [32] based his method on chart pattern analysis as the 2D picture by submitting picture data as features to a feed-forward neural network and training it to recognize the winning geometric patterns. Svitlana Galeshchuk [34] used macroeconomic factors such as (GDP growth, unemployment, wages), current and capital account(current account balance, openness as ratio of total import and export to GDP), public and private foreign debt, capital flows, and the ratio of international reserves to 3 months import, global variables (interest rates and price ratios), as input features to a deep neural network. This method is focused mainly on a novel feature based on currency clusters, but it's also useful for emerging currencies markets. This approach is based on Autoregressive Integrated Moving Average (ARIMA) models and Exponential Smoothing (ETS) models as times series models for its analysis and Stacked Long Short-term Memory for features processing and future stock prices prediction. Iman. Zabbah [35] enhanced an automated strategy based on EMA crossover, with 2 EMA lines of 2 different periods in a short time frame. This strategy is also based on Multilayer perceptron with linear scaling between the minimum and maximum values, as input features, to avoid false signal that triggers losing orders and distinguish winning EMA crossovers.

Nathan D'Lima [36] also chooses the same approach as Svitlana Galeshchuk [33] by basing their analysis on the hypothesis that time-series behavior is influenced by economic, political, and psychological factors. And he proposed a solution that combines a Multi-layer perceptron (MLP) with back-propagation and Adaptive Neuro-Fuzzy inference System (ANFIS). First, the MLP takes as inputs the previous day closing prices, simple moving average (SMA), exponential moving average (EMA), and Rate of Change (ROC) as momentum oscillators. And the ANFIS takes the previous day's closing prices and SMA. Then Mean Square Error (MSE) and Mean Absolute Error (MAE) for performance metrics.

On the other hand, Zhixi Li and Vincent Tam [10] found research on momentum and reversal effects unconvincing. This is because they cannot be applied as a trading strategy on a real-world investment portfolio, and the reason was the lack of predictive ability. So, they make a deep comparison between Decision Tree (D.T.), Support Vector Machine (SVM), Multilayer Perceptron Neural Network (MLP), and Long Short-Term Memory Neural Network. For them, SVM can be a source of profitable trading strategies.

D. Neural Network-based Approaches

In the same radius, we can mention Svetlana Borovkova and Ioannis Tsiamas proposed a group of Neural Networks of Long-Short-Term Memory (LSTM) for intraday trading based on technical pre-scan analysis [37]. This approach is quite good, but it does not consider the functional parameters: Momentum reversal cases, massive trend following, and volatility of the chosen market, which is the basis of our motivation. This technique works by weighting individual models in proportion to their past performance, permitting them to deal with possible non-stationarities by using a new approach. The sector then measures model performance under the operational characteristics of the receiver. Finally, they evaluate the predictive power of their model on several high-

value U.S. stock prices-cap and compare it to Lasso and Ridge. The proposed model was better than either the reference model or the equally weighted reference model sets. But this will not stop us from comparing the approach to the LSTM-based methods.

Seyed Taghi Akhavan Niaki and Saeid Hoseinzade study show the prediction of the daily direction of (S & P 500) index using an ANN. The main purpose is to select the most influential (features) characteristics of the proposed ANN that affect the day-to-day direction of S& P 500 (response). Experiments were conducted to find correct parts among 27 potential variables, projected as input nodes of the trained ANN. Using this methodology, the results of this study, using the most accurate features, can predict the daily S&P 500, which is much better than the existing logic model. In addition, the experimental results from the use of studied ANN on trial trading could significantly increase trading benefits by using a buy-and-hold strategy [38]. This has the same empiric context as our approach but does not fit our model for one reason: market behavior. This technique can be effective on trend following or momentum reversal but is not adaptable to volatile markets; the S&P 500 is a long-term trading index.

In their paper, Bang Xiang Yong, Mohd Rozaini Abdul Rahim, and Ahmad Shahidan Abdullah have proposed a trading system using a deep neural network (DNN) to predict the Singapore Stock Exchange Index (SESI) prices. Using the FTSE Straits Time Index (STI) in the coming days, test it through market simulations on historical daily prices [39]. The DNN has 40 nodes as the input layer, the last ten days, opening, closing, minimum and maximum price, and consists of 3 hidden layers with ten neurons per layer. It is trained by using stochastic gradient descent with backward propagation performed by multicore processing. They evaluated this approach using RME and MAPE and found that the profit factor was up 70%. It is not recommended to compare this method with ours because the latter is somewhat long-term, whereas our approach is intraday and mainly looks at the volatility phenomena. But we are going to be inspired by this method because it is based on deep learning.

Shuanglong Liu chooses the combination (CNN-LSTM) convolutional neural network and long-term memory, Chao Zhang, Jinwen Ma, to produce a model and analyze quantitative selection in stock markets [40]. First, the CNN method is used to establish the quantitative strategy of stock selection and then the quantitative approach of synchronization to improve profits using the LSTM. Their experiments show that the CNN neural network model can be successfully applied in developing a quantitative strategy and producing better yields than the benchmark. This approach is quite far from our study and cannot be compared with our practice due to its use of quantitative selection, but still one of the approaches that strands among state-of-art.

In their study, Wing Ki Liu and Mike K. P. incorporated the GARCH method on a neural network recurrent to model the price index's volatility on the S&P 500 financial market [41]. They produced a hybrid LSTM-GARCH model because it can consider the potential non-linearity of volatility at any time (t) and then predict the future volatility of the target index. This model is very similar to one of the channels of our CMLP, which led us to integrate it as a modal on an N.N. architecture for comparison.

On another radius, precisely on the study of the trend, G. Peter Zhang has set up a hybrid model ARIMA-LSTM, which will train an RNN based on three main features, the trend $\boldsymbol{m_t}$, the seasonal component $\boldsymbol{s_t}$ and the noise $\boldsymbol{\epsilon_t}$. This model represented the 2nd modal on the N.N. architecture of comparison [42].

E. Similar N.N. Architecture: MultiModal Learning based

Our case study is near Multi-Modal Deep Learning-based studies, which involve data from different input models by combining distinct modalities or groups of information with evident quantitative influence over the prediction output.

Our world considers several modalities, and a modality is a dataset about a lived experience in a specific context in real life. Most research focuses on sensory modalities representing the classic case, mainly based on natural language, visual, and vocal signals. It's the best approach to help artificial intelligence understand the world around us and interpret multi-modal messages from each modality. The classic case, which uses Multi-modal learning, has five main challenges:

- Representation: Summarization of multi-modal data (Heterogeneity)
- Translation: Mapping and Relationships between modalities (Open-Ended or Subjective)
- Alignment: Mapping and Relationships between sub-elements
- Fusion: Joining two or more modalities
- Co-learning: Knowledge transfer between modalities

There are two categories of multi-modal representations: joint and coordinated [43]. The collaborative approach consolidates unimodal signals into one unique model, shown in Fig. 5, and following this equation:

$$x_m = \mathrm{f}(x_1, \ldots, x_n)$$

Whereas coordinated handle unimodal signals separately, as shown in Fig. 6 and the following equation:

$$\mathrm{f}(x_1) \sim \mathrm{g}(x_2)$$

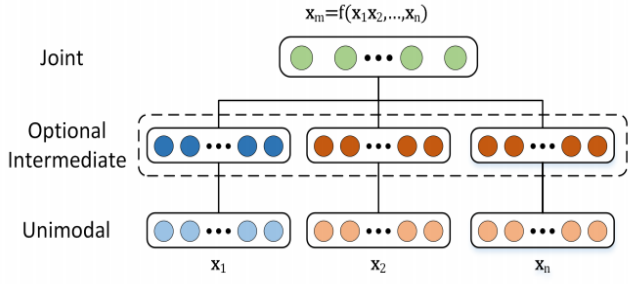


Fig. 5. Multi-Modal Joint Representation.

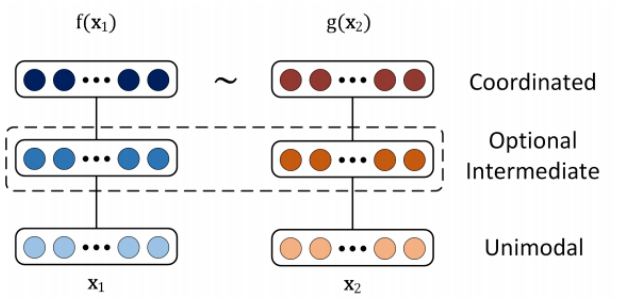


Fig. 6. Multi-Modal Coordinated Representation.

The generalized autoregressive conditional heteroskedasticity (GARCH) approach is used in financial markets volatility estimation. This technic is combined with intrinsic mode functions derived from empirical mode decomposition as a hybrid neural network [44]. They proposed an intelligent system that has the same properties as a multi-modal learning approach.

F. CSP, Share Memory by Communicating and Channeling

Communication is an exchange of specific data or explicit messages between specific N processes (explicit messages are used in distributed systems against shared memory systems). Messages are sent in two different ways, the asynchronous way that allows synchronization between emitter and receiver. The emitter process must wait for its ACK (acknowledgment) message to be received [45], [46]. And asynchronous way where the emitter does not wait for ACK and hand over to next process through shared variables which each process had public access to (used by standard memory systems) or by RPC (Remote Procedure Call). Synchronization links one process to another process and performs the exchange of control information between processes. In other words, if a process needs data produced by another process and this specific data is not available, the process must wait until it's available. The big challenge in process interleaving is their concurrency on data is that we must use an excellent technic to keep our system runnable. An interleaving of two sequences, s, and t, is a constructed U sequence from the events of s and t so that the events of s preserve their order in U and those of t. Hence, the birth of CSP is a method of communication and synchronization by using shared memory in multiprocessing. Instead of communicating by sharing memory (the traditional threading model used by famous languages like C and C++), based on sharing their data structures in memory, then locking used resources for a thread-safe mode-based communication between threads and mainly the processes. We will focus on the model-based approach used by Golang, which is sharing memory by communication. This concept is based on CSP [46] for its implementation using goroutines and channels as programming primitives. Goroutines are based on dynamic stack size, the reason why we called them "Green thread," the size of each dedicated segment on processes stack grows as needed depending on the data used by current tasks. Instead of the classic approach that uses a fixed stack size. This new technic allows us to create an infinite number of concurrent tasks. Goroutines start quickly instead of threads start-up and come with integrated primitives that allow them to communicate safely: Channels. There is Multi-Core Parallel Programming in Go demonstration [45], [47], [48] based on parallel integration for P.I. calculation and shows its performance by using a different number of parallel cores. We will build our trading architecture in charge of data retrieving, features extraction, and decision adaptation using a Goroutine-based implementation that fits our Channeling approach.

III. METHOD

Our approach is based on channeled multi-layer perceptron, in other words, building a feed-forward neural network that has a channel for each features group as follows:

A. Trend Features Group for T-Channel

Based on 3 EMAs (Exponential Moving Averages) on what we call an anchor time frame, in our case, a 1 Hour as long-term time frame. The main reason behind this is that there is less volatility on an H1 time frame and gives a safe deduction on market trend if it's in an uptrend or a downtrend. For trend determination, we use 3 EMA of different periods. Of course, we can use another period in our case (8, 13, 21), but those stay optimal for our approach. We also use the crossover to define the current hour trend, and the best method for this is 3 EMA crossover as cited in related works. After producing the 3 AMS data, we must select the qualified characteristics to be submitted to the trend channel. In this case, the growth rate of each EMA and the status ({1} for above - {-1} for below - {0} on the same line) between the first EMA (8 as a period) and the second (13 periods) the status between the second and third (21 periods), as shown in Table I:

TABLE I. SAMPLE OF TREND FEATURES GROUP FOR T-CHANNEL

Row	time	ema_1_gr	ema_2_gr	ema_3_gr	ema_1_2	ema_2_3
0	05:00	-1.23445	-0.7715	-0.54004	0	0
1	06:00	0.92575	0.46288	0.30858	-1	-1
2	07:00	0.69427	0.46286	0.30857	0	-1
3	08:00	3.16177	2.08242	1.31126	1	1
4	09:00	2.39004	1.6965	1.15686	1	1

B. Volatility Features Group for V-Channel

Based on the Bollinger Bands discovered by computing a 2-standard deviation of 20 periods SMA (Simple Moving Average) on an active trading frame, a 1 minute as a short time frame, where the volatility is considerably high. The standard deviation gives us a safe view of specific financial instrument volatility if it's stable or ultra-active. Technically, the standard deviation produces two bands, an upper and a lower band. If the upper band and lower band are far from the SMA, this means two things, the chosen financial instrument

is too volatile, and any crossing between the closing price and the upper or lower must be taken as trend reversal. After computing the 20-Period SMA and 2-Standard-Deviation, we choose the qualified features for the volatility channel. The distance between closing price and up or down band, and the distance between SMA and the up or down band, as shown on Table II:

TABLE II. SAMPLE OF VOLATILITY FEATURES GROUP FOR V-CHANNEL

Row	time	price_up_bb	price_low_bb	sma_up_bb
19	12:26	0.00129	0.00315	0.00222
20	12:27	0.001	0.00268	0.00184
21	12:28	0.00086	0.00226	0.00156
22	12:29	0.00065	0.00187	0.00126
23	12:30	0.00084	0.00124	0.00104

C. Momentum Features Group for M-Channel

Based on Stochastic RSI (Relative Strength Index) as cited in related works, a stochastic oscillator is applied on a set of RSI. The main goal behind using Stoch RSI is to detect the position reversal of a specific financial instrument by observing the crosses between RSI lines and oversold or overbought lines. We applied this on an active trading frame, in our case, a 1 minute as a short time frame. After computing the Stock RSI with {14, 14, 3, 3} as parameters, which are defined for our case. The distinguishing features for M-channel are the distance between RSI line 1 and oversold line, the distance between RSI line 1 and overbought line, the distance between RSI line 2 and oversold line and finally between RSI line 2 and overbought line, as shown on Table III:

TABLE III. SAMPLE OF MOMENTUM FEATURES GROUP FOR M-CHANNEL

Row	time	L14	H14	K_80	K_20	D_80	D_20
15	15:06	1.30374	1.30476	17.0	77.0	8.33	68.33
16	15:07	1.30378	1.30476	2.0	62.0	7.33	67.33
17	15:08	1.30378	1.30476	-3.0	57.0	5.33	65.33
18	15:09	1.30388	1.30476	5.0	65.0	1.33	61.33
19	15:10	1.30388	1.30476	12.0	72.0	4.67	64.67

D. Channeled MLP for TVM (Trend – Volatility – Momentum) Features Groups

As shown in Fig. 7, the main goal behind channeling a Multilayer Perceptron is to add a functional weight to each feature group and then parallelize their execution to enhance the speed of training and execution. In our case, we have three groups of features as follows:

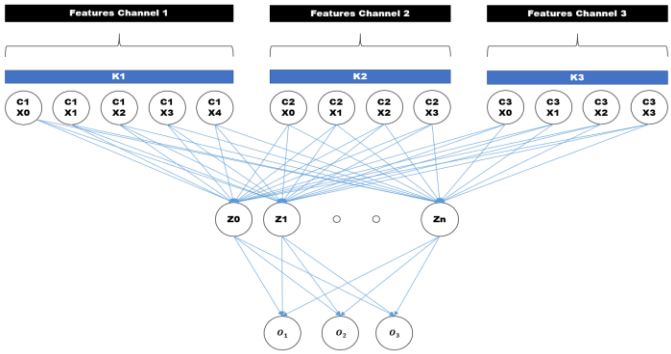


Fig. 7. Channeled Multi-Layer Perceptron Architecture Applied on Three Specific Channels.

- Trend group (Channel 1): composed of 5 features ema_1_gr, ema_2_gr, and ema_3_gr, which are the growing rate of EMAs then ema_1_2 and ema_2_3 which are respectively the EMA 1 / EMA 2 and EMA 2 / EMA 3 crossing status. We assign K1 to this channel as the initial weight.
- Volatility group (Channel 2): composed of 4 features price_up_bb, price_low_bb, sma_up_bb, and sma_low_bb, which are respectively the distance between price and upper band, the distance between price and lower band, the distance between SMA and upper band, and the distance between SMA and lower band. We assign K2 to this channel as the initial weight.
- Momentum group (Channel 3): composed of 4 features K_80, K_20 which are RSI line 1 and oversold line, the distance between RSI line 1 and overbought line, then D_80 and D_20, which are the distance between RSI line 2 and oversold line and finally between RSI line 2 and overbought line. We assign K3 to this channel as the initial weight.

The expected outputs are O1, O2, and Q3, respectively Buy signal, Sell signal and Wait for Signal. These outputs result from expected historical profits in the past. If the max profit is greater than 1, we consider it as a buy signal which has {1} as an expected feature. If the min profit is less than -1, we consider it as a sell signal which has {2} as the expected value. In the case where profit is between -1 and 1, the expected value is the wait signal which has {3} as the expected value.

E. Channeled MLP Execution (1 Iteration of Training)

Fig. 8 shows the execution of a Multi-layer perceptron using our approach, channeling a feed-forward neural network to parallelize each node's computation from the input layer to the output layer. It is considered that we have two channels which are composed of $(X_0, X_1, X_2, \dots, X_n)$ features, and are linked to each node of the hidden layer with $(W_0, W_1, W_2, \dots, W_n)$ weights. There are 2 (Depending on the number of channels) concurrent routines in charge of summation for each hidden layer node by respecting each channel weight. When summation of each channel is completed, a signal is triggered to start submission of sums to

activation function to produce (In the first iteration of the MLP training) Y + $\boldsymbol{\varepsilon}$ with Y as output and $\boldsymbol{\varepsilon}$ as an error. In our case, the used activation function is Sigmoid, and the first iteration is computed by using a random parameter for each node.

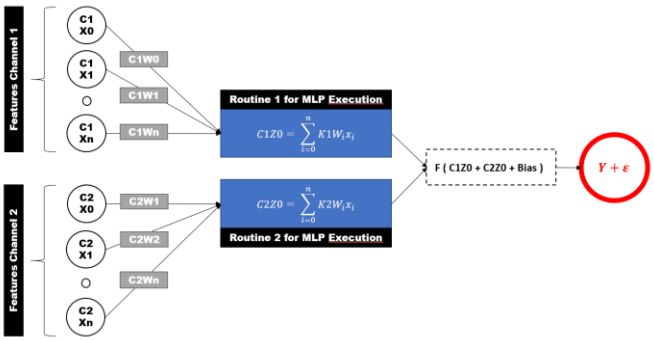


Fig. 8. Channeled MLP Executed by Channels Dedicated Routines.

F. Channelled MLP Back Propagation (Sigmoid Derivative Function)

In Fig. 9, we will see how back-propagation of produced error is done using our approach, which is based on channeling the propagation of error and applying it on the delta of each node. For each output node, there is a set of routines dedicated to each node of the hidden layer to compute the delta by using the derivative of the sigmoid action function, as shown in the Routine H_0 for Z_0 delta calculation. Then for each hidden layer node, there is a set of routines which is respectively dedicated to each node of the input layer to compute the delta by using the derivative of the sigmoid action function, as shown in the Routine I_0 and I_n for respectively X_0 and X_n Delta calculation. The process can be reproduced if there is more than one layer in the hidden layer. For concurrency synchronization, the computation of the previous layer must wait for the current layer computations to be completed.

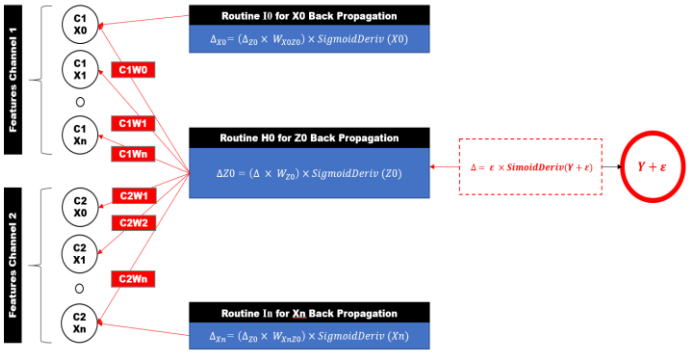


Fig. 9. Channeled Back-Propagation using Layer Set of Routines.

G. Channeled MLP Weights Adjustment (using Learning Rate)

Fig. 10 shows how weights adjustment is made for each node using our approach, based on channeling this process for each iteration. For each output node, there is a set of routines dedicated to nodes of the hidden layer to compute the learning rate, the delta, and the value of each node. We can see this in the Routine H_0 for Z_0 new weight calculation. Then for each input node, there is a set of routines that are dedicated to each node of the input layer to compute the learning rate, the delta, and the value of each node, it's represented by routine I_0 and I_n which are respectively in charge of W_{x_0} and W_{x_n} New weight calculation. The process can be reproduced if there is more than one layer in the hidden layer. However, for concurrency synchronization, the computation of the previous layer must wait for the current layer computations to be completed.

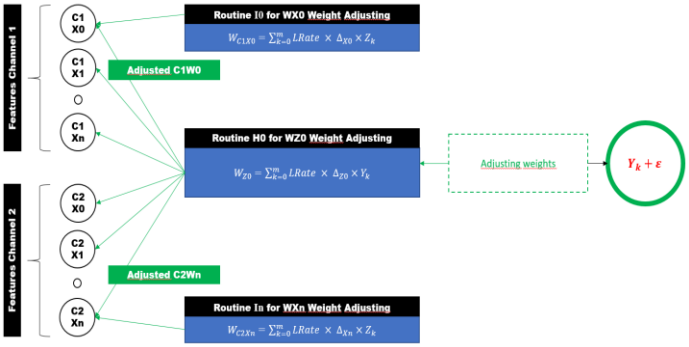


Fig. 10. Channeled Weights Adjustment.

H. Targeted Comparison Model: LSTM, GARCH, and ARIMA

To compare our model, as shown in Fig. 11, we will use a multi-modal neural network based on two different modals:

- The first modal is an LSTM ANN that will have input data from the GARCH approach, representing the study of volatility considering the historical returns of the preceding periods.
- The second modal is also an LSTM that will inject data from the ARIMA study that represents the trend of the target market.

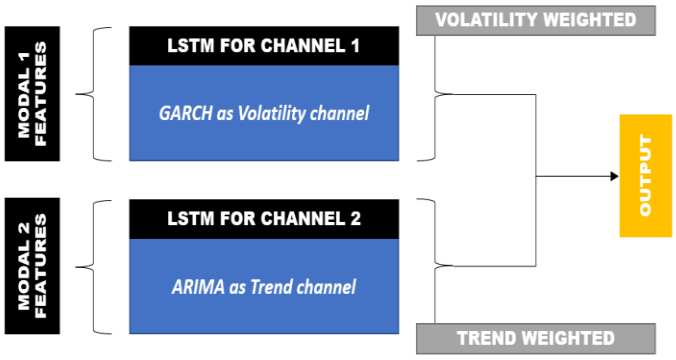


Fig. 11. GARCH-ARIMA MultiModal based on LSTM.

About the GARCH modal, this study follows a hybrid model: LSTM model and GARCH framework. The model first defines that the return $\boldsymbol{p_t}$ follows a standardized t distribution with conditional variance $\boldsymbol{h_t}$.. The conditional variance $\boldsymbol{h_t}$ is then modeled by LSTM. The objective function for the estimation of the parameter is the negative

loglikelihood. The GARCH model is conventionally used for profit modeling because it takes into account the dynamics nature of the conditional variance of performance. The GARCH(1,1) model can be expressed as:

$$p_t = \epsilon_t \sqrt{h_t} \quad | \quad \epsilon_t \sim N(0,1)$$

$$h_t = \alpha_0 + \alpha_1 h_{t-1} + \beta_1 p_{t-1}^2$$

$$\alpha_0 > 0 \quad | \quad \alpha_0, \alpha\beta_0 \geq 0$$

The stationary condition of the GARCH model (1,1) is:

$$\alpha_1 + \beta_1 < 1$$

In 2 first formula, the profit at time t, $\boldsymbol{p_t}$, follows a normal distribution with a zero mean and an $\boldsymbol{h_t}$variance, given the information up to time t-1. The variance at time t is the linear combination of conditional and square returns at time t1. The architecture of $\boldsymbol{h_t}$ is illustrated in Fig. 12. In the model, we will use the respective values of the previous **m** periods (x_{t-m}, ..., x_{t-1}) to predict the conditional variance of current h_t Returns.

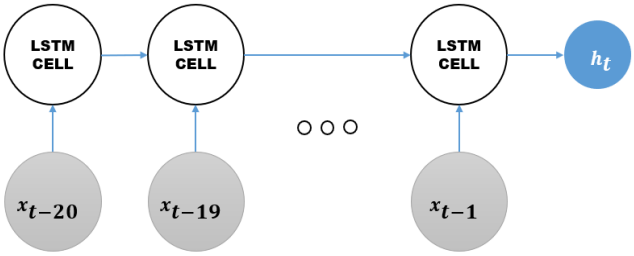


Fig. 12. GARCH based LSTM for Volatility.

Next comes the modal ARIMA/LSTM, a hybrid model where the study focuses mainly on the trend. This will be based on the variations of the triplet (m_t, s_t, ϵ_t) to predict m_t, this is a fairly simple model to consider the trend of the target market, specifically on forex, where the behavior does not always follow a uniform trend. The idea is to transform the original y_t Time-series into a smooth function by isolating it from its seasonal sound. To do this, we proceed to the first step of the Box and Jenkins method, which is $y_t = m_t + s_t + \epsilon_t$ Decomposition. Three functions are obtained: trend m_t, seasonal component and ϵ_t Noise. s_t and ϵ_t are kept to reproduce seasonality and put the forecast values into the confidence intervals using ϵ_t. From now on, m_t is smooth because it is free of fluctuation. Therefore, we can approach it through a network of neurons to know its dynamics and thus predict its future. Moreover, as it is a time series, the observations are sequential, so we use RNN of LSTM.

IV. Experiment

A. OANDA Broker – GBP/USD Instrument – REST API

For experimentation, we used OANDA as a broker due to available features for programmers. The most important feature for us is his REST API which provides access to historical data, price streaming, and orders management. The financial instrument used in this experiment is GBS/USD (Great Britain Pound against United States Dollar). The used OANDA API calls in our case are:

```
body=$(cat << EOF
{
"order": {
"units": {0},
"instrument": {1},
"timeInForce": {2},
"type": {3},
"positionFill": "DEFAULT"
}
} EOF
)
curl \
-X POST \
-H "Content-Type: application/json" \
-H "Authorization: Bearer <AUTHENTICATION TOKEN>" \
-d "$body" \
https://api-fxtrade.oanda.com/v3/accounts/<ACCOUNT>/orders
```

Where:

- {0}: Number of units to buy (Positive number) or sell (Negative number)
- {1}: Used financial instrument in our case GBP/USD
- {2}: In our case, we will use FOK (Fill Or Kill)
- {3}: The initial position which is set to default

B. Historical Data Retrieving then Features and Expected Values Production

This stage is dedicated to TVM (Trend-Volatility-Momentum) winning and losing patterns production. As a requests wrapper, we used a Golang based program, which is in charge of retrieving historical prices and computing them into usable features and expected values. This program takes the prepared features as a flat set of indicators for the first test, which is a normal MLP training, then taken as three groups (TVM) of features in the second test, which is the implementation of our approach, the channeled MLP training.

In the first stage, we must produce the T-CHANNEL (Trend Channel). In Table IV, we can see the growth rate and the status of each of the 3 EMA periods. Then profit max and min generated by each model. If the maximum profit is greater than {1.0}, we consider it as a buy signal that has {1} as an expected feature. Otherwise, if the minimum profit is less than -1, we consider it as a sales signal that has {2} as expected value:

TABLE IV. MAX AND MIN PROFIT THEN SIGNAL GENERATED FROM T-CHANNEL FEATURES

ema_1_gr	ema_2_gr	ema_3_gr	ema_1_2	ema_2_3	max_pr	Min_pr	Signal
-1.23445	-0.7715	-0.54004	-1	-1	1	-2	2
0.92575	0.46288	0.30858	-1	-1	0,3	-5	2
0.69427	0.46286	0.30857	1	-1	0,1	-1	2
3.16177	2.08242	1.31126	1	1	2	-0,1	1
2.39004	1.6965	1.15686	1	1	4	-0,1	1

In the second stage, we must produce the V-CHANNEL (Volatility Channel). In Table V, we see the calculated distance between prices or SMA up and low bands (Bollinger bands), then the max profit and the min profit generated in the case of trading on each pattern, if the max profit is greater than 1, we consider it as buy signal which has [1] as expected feature, if the min profit is less than -1, we consider it as sell signal which has [2] as the expected value. In the case where profit is between -1 and 1, the expected value is the wait signal which has [3] as the expected value.

In the third stage, we must produce the M-CHANNEL (Momentum Channel). In Table VI, we can see the calculated distance between K-Stochastic-RSI and D-Stochastic-RSI lines, and respectively 80% and 20% lines. Next, the max profit and min profit generated in the case of trading on each model. If the max profit is greater than 1, we consider it as a buy signal that has [1] as expected characteristic, if the min profit is less than - 1, we consider it as a sell signal that has [2] as expected value. Finally, if the profit is between -1 and 1, the expected value is the waiting signal that has [3] as expected value.

TABLE V. MAX AND MIN PROFIT THEN SIGNAL GENERATED FROM V-CHANNEL FEATURES

price_up_bb	price_low_bb	sma_up_bb	sma_low_bb	max_pr	min_pr	Signal
0.00129	0.00315	0.00222	0.00222	0,1	-0,1	3
0.001	0.00268	0.00184	0.00184	1	-0,4	1
0.00086	0.00226	0.00156	0.00156	0,1	-0,7	3
0.00065	0.00187	0.00126	0.00126	0,2	-0,3	3
0.00084	0.00124	0.00104	0.00104	0,7	-1	2

TABLE VI. MAX AND MIN PROFIT THEN SIGNAL GENERATED FROM M-CHANNEL FEATURES

K_80	K_20	D_80	D_20	max_pr	Min_pr	Signal
17.0	77.0	8.33	68.33	0,2	-3	2
2.0	62.0	7.33	67.33	0,1	-2	2
3.0	57.0	5.33	65.33	0,9	-0,5	3
5.0	65.0	1.33	61.33	0,7	-0,2	3
12.0	72.0	4.67	64.67	0,3	-0,01	3

1) Normal Multi-layer perceptron training

We used the current parameters for training our MLP:

- Global parameters (Using K-Fold training):
  - *a)* Learning Rate: 0.01
  - *b)* Accuracy percentage: 0.67
  - *c)* Shuffle: True
  - *d)* Epochs: 100/200/500
  - *e)* Folds: 2/3/5
- For trend features:
  - *a)* Granularity: H1 (1 hour)
  - *b)* Number of candle patterns: 500
  - *c)* Max/Min profit: 1 / -1
- For volatility features:
  - *a)* Granularity: M1 (1 minute)
  - *b)* Number of candle patterns: 500
  - *c)* Max/Min profit: 1 / -1
- For momentum features:
  - *a)* Granularity: M1 (1 minute)
  - *b)* Number of candle patterns: 500
  - *c)* Max/Min profit: 1 / -1

Table VII shows the list of features in the input layer:

TABLE VII. THE COMPLETE LIST OF FEATURES SUBMITTED TO NORMAL MLP

Feature name	Associated name	Node notation
Short Period EMA (EMA 1) growing rate	ema_1_gr	X0
Medium Period EMA (EMA 2) growing rate	ema_2_gr	X1
Long Period EMA (EMA 3) growing rate	ema_3_gr	X2
Common status between EMA 1 / EMA 2	ema_1_2	X3
Common status between EMA 2 / EMA 3	ema_2_3	X4
Distance between price and upper Bollinger band	price_up_bb	X5
Distance between price and lower Bollinger band	price_low_bb	X6
Distance between SMA and upper Bollinger band	sma_up_bb	X7
Distance between SMA and lower Bollinger band	sma_low_bb	X8
Distance between K-Stochastic-RSI line and 80%	K_80	X9
Distance between D-Stochastic-RSI line and 80%	D_80	X10
Distance between K-Stochastic-RSI line and 20%	K_20	X11
Distance between D-Stochastic-RSI line and 20%	D_20	X12

Table VIII shows the list of expected values (Classes) in the output layer:

TABLE VIII. EXPECTED VALUES ON THE OUTPUT LAYER

Expected output	Name of Expected output	Description
1	Buy signal	Triggered buy signal on expected max profit obtained
2	Sell signal	Triggered sell signal on expected max profit obtained
3	Wait for signal	Triggered wait signal on expected max profit not obtained

2) Channeled Multi-layer perceptron training:

We used the same parameters in normal MLP training, but we added another parameter which is the associated weight for each channel:

- For trend features: K1 as channel associated weight
- For volatility features: K2 as channel associated weight
- For momentum features: K3 as channel associated weight

The channeled features are distributed in Table IX, as follows:

TABLE IX. LIST OF THE FEATURES SUBMITTED TO CHANNELED MLP

Feature name	Associated name	Node notation	Channel	Initial weight
Short Period EMA (EMA 1) growing rate	ema_1_gr	C1X0	C1	K1
Medium Period EMA (EMA 2) growing rate	ema_2_gr	C1X1	C1	K1
Long Period EMA (EMA 3) growing rate	ema_3_gr	C1X2	C1	K1
Common status between EMA 1 / EMA 2	ema_1_2	C1X3	C1	K1
Common status between EMA 2 / EMA 3	ema_2_3	C1X4	C1	K1
Distance between price and upper Bollinger band	price_up_bb	C2X1	C2	K2
Distance between price and lower Bollinger band	price_low_bb	C2X2	C2	K2
Distance between SMA and upper Bollinger band	sma_up_bb	C2X3	C2	K2
Distance between SMA and lower Bollinger band	sma_low_bb	C2X4	C2	K2
Distance between K-Stochastic-RSI line and 80%	K_80	C3X0	C3	K3
Distance between D-Stochastic-RSI line and 80%	D_80	C3X1	C3	K3
Distance between K-Stochastic-RSI line and 20%	K_20	C3X2	C3	K3
Distance between D-Stochastic-RSI line and 20%	D_20	C3X3	C3	K3

3) Multi-Modal learning based on GARCH and ARIMA: As shown in Section 2, we decided to build a multimodal N.N. system based on generalized autoregressive conditional heteroscedasticity (GARCH) and integrated autoregressive moving average (ARIMA). This hybrid approach is already used in volatility prediction using the ARMA(R,M) and GARCH(p,q) models[48]. The first group or modality is the variance group representing volatility that belongs to GARCH, based on Gaussian pseudo-random numbers. This modality is composed of an average model, a volatility process and a distribution. The second modality that represents the trend group that represents the ARIMA model.

The most important part is the study on volatility. For the first GARCH-oriented LSTM intended to receive volatility-related data

- The features chosen as input are the elements having a studied correlation and which constitute the vector

$$x_t = \{o_t, g_t, l_t, c_t\}$$

where o, g, l, and c are in the same order open, high, low, and the close price at the time (t), then the parameter p_t^2 that will allow each LSTM node to model the h_t, based on conditional variance using the objective loglikelihood function.

- The exits of the LSTM are the following:

$$p_t = P_t - P_{t-1}$$

Where:

p_t is the profit at **t,** and

P_t is the price at **t**

4) Classical hybrid approach: Then combine Bollinger Band with Stochastic RSI and use it as a trading strategy on GBP/USD couple. Intraday traders widely use this technic to make an exhaustive comparison to our strategy and see the benefits of using a Channeled MLP as a validation layer on a switching pattern scenario to be adaptive to any kind of markets behavior.

V. RESULTS

The back-testing execution is done on a machine with Ubuntu 18.04 as an operating system and Docker containers for worker nodes to provide a simulated aspect of distributed computing. All of this is running on a physical machine with an Intel i5 2.2 GHz 4 CORE processor and 8 G.B. of RAM.

For our test case, we have retrieved GBP/USD data of 2-time frames, 1-hour and 1-minute prices, then calculated their 3 EMA crossovers, Bollinger Bands, Stochastic RSI, GARCH, and ARIMA as shown in the method section, to produce the five needed datasets. Then came the normalization process to ensure the quality and performance of training. We then randomly split data into two datasets, 70% for training against 30% for test.

In Fig. 13, we see the scores reached by implementing normal Multilayer Perceptron, the multi-modal GARCH-ARIMA, and channeled MLP. It shows the scores for 2, 3, and 5 folds for 100, 200, and 500 epochs as training iterations. Due to dedicated channel weights, the Channeled Multilayer Perceptron has better scores than normal MLP and MM GARCH ARIMA.

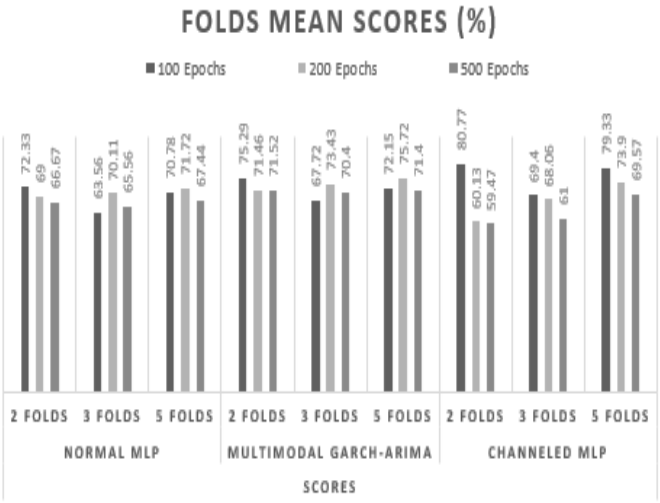


Fig. 13. Back-Testing Scores reached Comparison between Normal MLP, MM GARCH-ARIMA and our Channeled MLP.

In Fig. 14, we see the training time taken by implementing normal Multilayer Perceptron and channeled MLP. It shows the scores for 2, 3, and 5 folds for 100, 200, and 500 epochs as training iterations. Due to parallelized and concurrent node calculations, the Channeled Multilayer Perceptron has been trained faster than normal MLP and MM GARCH ARIMA.

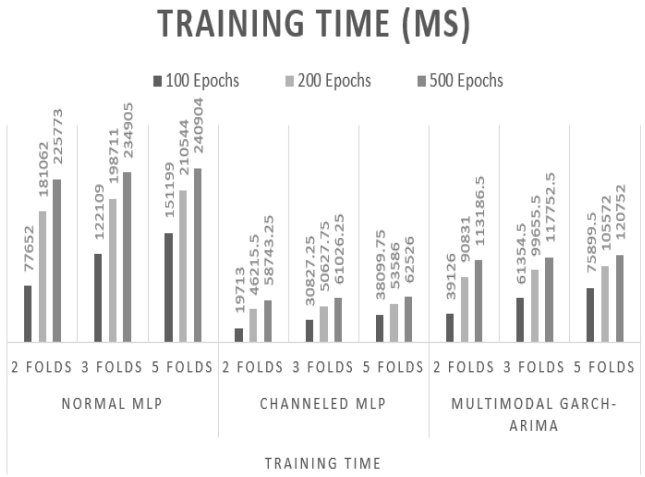


Fig. 14. Training Time Comparison between Normal MLP, MM GARCH-ARIMA and our Channeled MLP.

In Fig. 15, we see the execution time taken by implementing normal Multilayer Perceptron and channeled MLP. It shows the scores for 2, 3, and 5 folds for 100, 200, and 500 epochs as training iterations. The Channeled Multilayer Perceptron computes the given pattern and provides the associated output faster than normal MLP and MM GARCH ARIMA due to parallelized and concurrent node calculations.

In Table X, we observe the trades opened by normal Multilayer Perceptron, channeled MLP, and BB/SRSI, in the same day to see the efficiency of Channeled MLP decisions. The Channeled Multilayer Perceptron has opened six trades where 5 of them were positive with 9.7 cumulative earnings and a profit factor that reached 83%. The normal MLP has opened six trades where 3 of them had positive profits up to 2.15 as net profit. Then the MM GARCH-ARIMA opened six trades where 1 of them was positive with fatal negative profit (-14.3). The main problem was the lack of features that stands for momentum reversal in the market which does a lot of brutal reversal in its trend. And finally, BOLLINGER BAND / Stochastic RSI with six open trades where 1 of them was positive having 0.7 profit. Thus, we can say that channeled MLP made good decisions in opening market switching by trend, volatility, and momentum reversal in GBP/USD market.

We chose the MultiModal GARCH-ARIMA LSTM model to create a more appropriate and similar comparison context to our approach and thus target several existing methods cited in the state of the art, some of which are not directly comparable since the majority operate in different markets and apply to long-term investment types. In contrast, our approach targets those that affect short-term investment. Furthermore, our CMLP approach considers three indicators and assigns dynamic weights that vary according to the market behavior, which allows it to adapt to so-called aggressive markets. Using the other approaches in more volatile markets, the system triggers trades on necessary signals, hence its strength about the other market, which appears on the results obtained.

Fig. 15. Execution Time Comparison between Normal MLP, MM GARCH-ARIMA and our Channeled MLP.

TABLE X. COMPARISON OF PROFITS EARNED BETWEEN NORMAL MLP, MM GARCH-ARIMA AND OUR CHANNELED MLP

NN TYPE	Type	Market	Units	Profit (USD)	Half Spread Cost	Date
NORMAL MLP	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-1.45**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**2.3**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**2.8**	-1.8	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-2.4**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-2.3**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**3.2**	-1.7	15/06/2021
CHANNELED MLP	Buy Market	GBP/USD	20	**0**	-1.7	15/06/2021
	Close Trade	GBP/USD	20	**2.6**	-1.7	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.7	15/06/2021
	Close Trade	GBP/USD	20	**-4.6**	-1.7	15/06/2021
	Sell Market	GBP/USD	20	**0**	-1.9	15/06/2021
	Close Trade	GBP/USD	20	**2.5**	-2	15/06/2021
	Sell Market	GBP/USD	20	**0**	-2	15/06/2021
	Close Trade	GBP/USD	20	**2.4**	-1.9	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.4	15/06/2021
	Close Trade	GBP/USD	20	**5.6**	-1.4	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.4	15/06/2021
	Close Trade	GBP/USD	20	**1.2**	-1.8	15/06/2021
MultiModal GARCH-ARIMA	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-2.4**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-4.6**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**4**	-1.5	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.5	15/06/2021
	Close Trade	GBP/USD	20	**-5**	-1.5	15/06/2021
	Sell Market	GBP/USD	20	**0**	-1.5	15/06/2021
	Close Trade	GBP/USD	20	**-3.1**	-1.6	15/06/2021
	Sell Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-3.2**	-1.6	15/06/2021
BoLLINGER BAND / StoChastic RSI	Sell Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**10**	-1.7	15/06/2021
	Sell Market	GBP/USD	20	**0**	-1.8	15/06/2021
	Close Trade	GBP/USD	20	**1.8**	-1.6	15/06/2021
	Sell Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-2**	-1.7	15/06/2021
	Sell Market	GBP/USD	20	**0**	-1.7	15/06/2021
	Close Trade	GBP/USD	20	**-2.3**	-1.8	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.8	15/06/2021
	Close Trade	GBP/USD	20	**-3**	-1.6	15/06/2021
	Buy Market	GBP/USD	20	**0**	-1.6	15/06/2021
	Close Trade	GBP/USD	20	**-3.8**	-1.8	15/06/2021

VI. Discussion

Most of the methods cited on the state of the art are methods used for algo-trading based on a single indicator and are not based on neural networks, while we target multiple and different type indicators to allow for varying market knowledge. But even we thought it useful to quote them. The only strategy that can make target method topic for our comparison is the neural multi-modal network architecture, only because this technique can take features of several types and group them by observation nature. In our study we shed more light on the approaches that use the multi modal base on 2 LSTM, the first is the LSTM-GARCH which takes into account features oriented towards volatility. And the second is the LSTM-ARIMA which takes into account the market trend, and their weights are static so do not allow to act on markets that are volatile and – or trending. Our approach takes into account another parameter which is the reversal momentum (Stochastic RSI) as oscillator. And puts more weight on the group that characterizes the market has a given period (e.g.: Gives more weight to Group V - volatility if the market is volatile or-and more weight to Group T - trend if the market has a strong trend). The back-testing of our trading strategy (Method) is based on training an NN to recognize a false signal and a real signal or a neutral signal of waiting for opportunity. The scores obtained and displayed in Fig. 11 mean that our CMLP approach is more accurate than the MM LSTM GARCH-ARIMA. The results of our experiments are more focused on trades having positive incomes which represent the main criteria of efficiency. For training / testing we used a proportion of 70% - 30% to back test our approach. We made a focus on sentiment analysis on work. In this paper we bring more light on the statistical aspect for prediction. We have already brought attention on cryptocurrencies that are influenced by communities in Twitter.

VII. Conclusion

The forex market is the more volatile market among financial markets. Trading on it at high frequency is so complex and needs a well-designed algorithm, the reason why we use machine learning, especially the Multi-layer perceptron, which is the best approach for time series. The faced challenges are training speed and execution, more than that, the accuracy of taken decision on open or not open a trade, and if the decision is opening a trade, should the MLP choose to buy, sell or wait. Channeling features helps us a lot with this. However, our approach is still technically crud and needs more ameliorations for some minor details, to be mature and well prepared for aggressive markets, to avoid bad decisions that can make us lose more trades than expected. As future works we are planned to make a focus on sentiment analysis as category of features selection in our current work. We have already brought attention on cryptocurrencies that are influenced by communities in Twitter; this area is also a subject of volatility research.

==========

Declaration of interests

☒ ***The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.***

☐ *The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:*

==========

References

[1] O. Chantarakasemchit, S. Nuchitprasitchai, and Y. Nilsiam, "Forex Rates Prediction on EUR/USD with Simple Moving Average Technique and Financial Factors," in 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Jun. 2020, pp. 771–774. doi: 10.1109/ECTI-CON49241.2020.9157907.

[2] V. Patil, N. Somani, A. Tadvi, and V. Attar, "Algorithmic Forex Trading using Combination of Numeric Time Series and News Analysis," in 2018 4th International Conference for Convergence in Technology (I2CT), Oct. 2018, pp. 1–5. doi: 10.1109/I2CT42659.2018.9058285.

[3] Ji Hoon Jang et al., "Accelerating forex trading system through transaction log compression," in 2014 International SoC Design Conference (ISOCC), Nov. 2014, pp. 74–75. doi: 10.1109/ISOCC.2014.7087602.

[4] T. C. O. Leslie, C. L. N. David, and Y. Lee, "Prediction of Forex trend movement using linear regression line, two-stage of multi-layer perceptron and dynamic time warping algorithms," 2016. doi: 10.32890/jict2016.15.2.6.

[5] A. R. P. Putra, A. E. Permanasari, and S. Fauziati, "I forex trend prediction technique using multiple indicators and multiple pairs correlations DSS: A software design," in 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Oct. 2016, pp. 1–5. doi: 10.1109/ICITEED.2016.7863248.

[6] A. Kale, O. Khanvilkar, H. Jivani, P. Kumkar, I. Madan, and T. Sarode, "Forecasting Indian Stock Market Using Artificial Neural Networks," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Aug. 2018, pp. 1–5. doi: 10.1109/ICCUBEA.2018.8697724.

[7] A. Noertjahyana, A. Noertjahyana, Z. A. Abas, and Z. I. M. Yusoh, "Combination of Candlestick Pattern and Stochastic to Detect Trend Reversal in Forex Market," in 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Dec. 2019, pp. 1–4. doi: 10.1109/TIMES-iCON47539.2019.9024485.

[8] Y. Zhai, A. Hsu, and S. K. Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," in Advances in Neural Networks – ISNN 2007, Berlin, Heidelberg, 2007, pp. 1087–1096. doi: 10.1007/978-3-540-72395-0_132.

[9] Y. Alperovich, M. Alperovich, and A. Spiro, "Forecasting Price Trends in Financial Markets," in 2018 Eleventh International Conference "Management of large-scale system development" (MLSD, Oct. 2018, pp. 1–3. doi: 10.1109/MLSD.2018.8551778.

[10] Z. Li and V. Tam, "A Machine Learning View on Momentum and Reversal Trading," p. 16, 2018.

[11] T. J. Moskowitz, Y. H. Ooi, and L. H. Pedersen, "Time series momentum," Journal of Financial Economics, vol. 104, no. 2, pp. 228–250, May 2012, doi: 10.1016/j.jfineco.2011.11.003.

[12] "Moving Average Envelopes and Bollinger Bands," in Technical Analysis and Chart Interpretations, John Wiley & Sons, Ltd, 2016, pp. 215–221. doi: 10.1002/9781119204800.ch17.

[13] M. Ghorbani and E. K. P. Chong, "Stock price prediction using principal components," PLOS ONE, vol. 15, no. 3, p. e0230124, Mar. 2020, doi: 10.1371/journal.pone.0230124.

[14] H. He, J. Chen, H. Jin, and S.-H. Chen, "Trading Strategies Based on K-means Clustering and Regression Models," in Computational Intelligence in Economics and Finance: Volume II, S.-H. Chen, P. P. Wang, and T.-W. Kuo, Eds. Berlin, Heidelberg: Springer, 2007, pp. 123–134. doi: 10.1007/978-3-540-72821-4_7.

[15] S. Hansun and M. B. Kristanda, "Performance analysis of conventional moving average methods in forex forecasting," in 2017 International Conference on Smart Cities, Automation Intelligent Computing Systems (ICON-SONICS), Nov. 2017, pp. 11–17. doi: 10.1109/ICON-SONICS.2017.8267814.

[16] C. Xiao, W. Xia, and J. Jiang, "Stock price forecast based on combined model of ARI-MA-LS-SVM," Neural Comput & Applic, vol. 32, no. 10, pp. 5379–5388, May 2020, doi: 10.1007/s00521-019-04698-5.

[17] S. Hansun, "FX forecasting using B-WEMA: Variant of Brown's Double Exponential Smoothing," in 2016 International Conference on Informatics and Computing (ICIC), Oct. 2016, pp. 262–266. doi: 10.1109/IAC.2016.7905726.

[18] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[19] S. Lauguico et al., "A Fuzzy Logic-Based Stock Market Trading Algorithm Using Bollinger Bands," in 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Nov. 2019, pp. 1–6. doi: 10.1109/HNICEM48295.2019.9072734.

[20] M. Butler and D. Kazakov, "Particle Swarm Optimization of Bollinger Bands," in Swarm Intelligence, Berlin, Heidelberg, 2010, pp. 504–511. doi: 10.1007/978-3-642-15461-4_50.

[21] T. W. A. Khairi, R. M. Zaki, and W. A. Mahmood, "Stock Price Prediction using Technical, Fundamental and News based Approach," in 2019 2nd Scientific Conference of Computer Sciences (SCCS), Mar. 2019, pp. 177–181. doi: 10.1109/SCCS.2019.8852599.

[22] T. Chong and W.-K. Ng, "Technical analysis and the London stock exchange: Testing the MACD and RSI rules using the FT30," Applied Economics Letters, vol. 15, pp. 1111–1114, Feb. 2008, doi: 10.1080/13504850600993598.

[23] A. Rodríguez-González et al., "Improving Trading Systems Using the RSI Financial Indicator and Neural Networks," in Knowledge Management and Acquisition for Smart Systems and Services, Berlin, Heidelberg, 2010, pp. 27–37. doi: 10.1007/978-3-642-15037-1_3.

[24] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights Imaging, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[25] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," Neural Computation, vol. 29, no. 9, pp. 2352–2449, Jun. 2017, doi: 10.1162/neco_a_00990.

[26] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural network and its application in image classification: a survey," OE, vol. 58, no. 4, p. 040901, Apr. 2019, doi: 10.1117/1.OE.58.4.040901.

[27] T. Das and G. Saha, "Addressing big data issues using RNN based techniques," Journal of Information and Optimization Sciences, vol. 40, no. 8, pp. 1773–1785, Nov. 2019, doi: 10.1080/02522667.2019.1703268.

[28] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to Construct Deep Recurrent Neural Networks," arXiv:1312.6026 [cs, stat], Apr. 2014, Accessed: Nov. 23, 2020. [Online]. Available: http://arxiv.org/abs/1312.6026.

[29] R. Bai, J. Zhao, D. Li, X. Lv, Q. Wang, and B. Zhu, "RNN-based demand awareness in smart library using CRFID," China Communications, vol. 17, no. 5, pp. 284–294, May 2020, doi: 10.23919/JCC.2020.05.021.

[30] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron: Architecture Optimization and training with mixed activation functions," in Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, New York, NY, USA, Mar. 2017, pp. 1–6. doi: 10.1145/3090354.3090427.

[31] M. Aitkin and R. Foxall, "Statistical modelling of artificial neural networks using the multi-layer perceptron," Statistics and Computing, vol. 13, no. 3, pp. 227–239, Aug. 2003, doi: 10.1023/A:1024218716736.

[32] G. I. Sher, "Evolving Chart Pattern Sensitive Neural Network Based Forex Trading Agents," arXiv:1111.5892 [cs], Jan. 2012, Accessed: Nov. 23, 2020. [Online]. Available: http://arxiv.org/abs/1111.5892

[33] O. Fathi, "Time series forecasting using a hybrid ARIMA and LSTM model," p. 7.

[34] S. Galeshchuk, "Neural networks performance in exchange rate prediction," Neurocomputing, vol. 172, pp. 446–452, Jan. 2016, doi: 10.1016/j.neucom.2015.03.100.

[35] I. Zabbah and E. Partovi, "Enhancing an Automated Trading Strategy Using Artificial Neural Networks," p. 4, 2012.

[36] N. D'Lima and S. Khan, "FOREX rate prediction using a Hybrid System," International Journal of Engineering Research and Technology, vol. 3, Nov. 2014.

[37] S. Borovkova and I. Tsiamas, "An ensemble of LSTM neural networks for high-frequency stock market classification," J FORECASTING, vol. 38, no. 6, pp. 600–619, 2019, doi: 10.1002/for.2585.

[38] S. T. A. Niaki and S. Hoseinzade, "Forecasting S&P 500 index using artificial neural networks and design of experiments," J Ind Eng Int, vol. 9, no. 1, p. 1, Feb. 2013, doi: 10.1186/2251-712X-9-1.

[39] B. X. Yong, M. R. Abdul Rahim, and A. S. Abdullah, "A Stock Market Trading System Using Deep Neural Network," in Modeling, Design and Simulation of Systems, Singapore, 2017, pp. 356–364. doi: 10.1007/978-981-10-6463-0_31.

[40] S. Liu, C. Zhang, and J. Ma, "CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets," in Neural Information Processing, Cham, 2017, pp. 198–206. doi: 10.1007/978-3-319-70096-0_21.

[41] W. Liu and M. So, "A GARCH Model with Artificial Neural Networks," Information, vol. 11, p. 489, Oct. 2020, doi: 10.3390/info11100489.

[42] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.

[43] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," arXiv:1705.09406 [cs], Aug. 2017, Accessed: Jul. 02, 2021. [Online]. Available: http://arxiv.org/abs/1705.09406.

[44] H. A. Pathberiya, C. D. Tilakaratne, and L. L. Hansen, "An intelligent system for forex trading: Hybrid ANN with GARCH and intrinsic mode functions," in 2017 Intelligent Systems Conference (IntelliSys), Sep. 2017, pp. 436–445. doi: 10.1109/IntelliSys.2017.8324331.

[45] R. M. Fujimoto and H. Feng, "A shared memory algorithm and proof for the generalized alternative construct in CSP," Int J Parallel Prog, vol. 16, no. 3, pp. 215–241, Jun. 1987, doi: 10.1007/BF01407934.

[46] C. A. R. Hoare, "Communicating sequential processes," Commun. ACM, vol. 21, no. 8, pp. 666–677, Aug. 1978, doi: 10.1145/359576.359585.

[47] N. Fikri, M. Rida, N. Abghour, K. Moussaid, and A. El Omri, "A Near Metal Platform for Intensive Big Data Processing Using A Novel Approach: Persistent Distributed Channels," in Proceedings of the 2020 5th International Conference on Big Data and Computing, New York, NY, USA, May 2020, pp. 1–5. doi: 10.1145/3404687.3404692.

[48] A. Prell and T. Rauber, "Go's Concurrency Constructs on the SCC," p. 7, 2012.

A Novel Cyber-attack Leads Prediction System using Cascaded R2CNN Model

P. Shanmuga Prabha[1]
Assistant Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

S. Magesh Kumar[2]
Professor, Department of Computer Science and Engineering Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University Chennai, Tamilnadu, India, Pincode: 602105

***Abstract*—Novel prediction systems are required in almost all internet-connected platforms to safeguard the user information to get hacked by intermediate peoples. Finding the real impacted factors associated with the Cyber-attack probes are being considered for research. The proposed methodology is derived from various literature studies that motivated to find the unique prediction model that shows improved accuracy and performance. The proposed model is represented as R2CNN that acts as the cascaded combination of Gradient boosted regression detector with recurrent convolution neural network for pattern prediction. The given input data is the collection of various applications engaged with the wireless sensor nodes in a smart city. Each user connected with a certain number of applications that access the authorization of the device owner. The dataset comprises device information, the number of connectivity, device type, simulation time, connectivity duration, etc. The proposed R2CNN extracts the features of the dataset and forms a feature mapping that related to the parameter being focused on. The features are tested for correlation with the trained dataset and evaluate the early prediction of Cyber-attacks in the massive connected IoT devices.**

Keywords*—Cyber security in smart devices; cyber security; cyber-attacks; internet of things; IoT devices; machine learning; wireless sensor networks

I. Introduction

In the current fast growing world, the demand on cyber security to users and IOT devices becomes mandatory due to the equivalent increase of cyber-attacks. The awesome growth of internet and networking models enables the users to get access the internet hassle free with flexible applications. Numerous cyber-attacks are predicted by the software and intrusion detection systems installed in the IOT devices and user accessible devices. Some of the frequently occurring cyber-attacks are denial-of-service(DoS) issue, software malwares, hacking due to unauthorized access, man in middle problem etc. the data industry get affected a lot due to the loss of secretive information that should be maintained [1].

Cyber security system is comprises of layers of protection starting from the physical layer to the application layer. Normally application layer resemble the end user connected with the certain network. Physical layer security plays a vital role in holding the data and making the secure communication with the data link layer. Nowadays in most of the systems, a powerful intrusion detection model is developed and installed. These detectors filter out the malwares to some extent. Working on achieving accurate intrusion detection systems are also focused by research team. The author in [1] discussed recently with intrusion detection system using tree method that is named as intruDTree that keeps the decision tree as base model. The implemented intruDtree creates various test cases and expressions to check the real impact of the action takes place in to the IOT device. [2]Evaluated a signature based approach in the development of intrusion detection system and intrusion prevention system. The malicious software damages the protection systems, in the form of worms that uses the same protocol to carry over the network. Machine learning based signature model is evaluated to adopt the changing frameworks and authenticate the new user every time. The level of authentication happens in each stages of network connectivity [3].

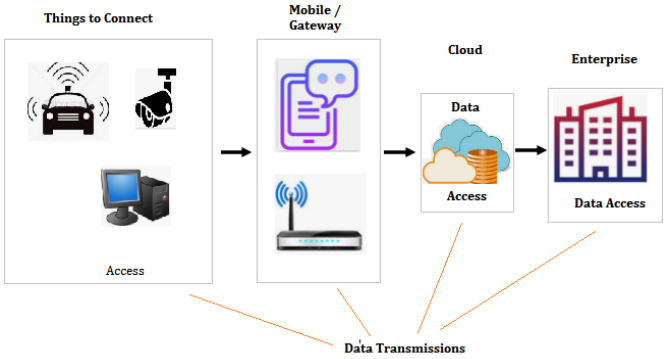


Fig. 1. Model of Cyber Security in IoT Devices.

Fig. 1 shows the architecture of cyber security framework. The model is explained in three phases. The first one is the data gathering phase in which the information in the form of images, videos, accessibility, unique codes, vehicle data and emergency information are collected. The second phase is the gateway or communication medium where the mobile gateway, data processing modules, android devices or wireless modems are connected. The cloud based accessibility also comes under the same category. The third phase is the application model in which the data is applied to the utilization of operations and control [4]. IoT based Cyber security monitoring for Internet dependent drones that gathers the tracking data and input patterns. Based on the available data, using Naïve Bayes model the system is able to identify

the cyber threats [4], further by addressing the existing issue of assumptions on information received need to be improving by utilization of better algorithms. Cyber security attacks get injected into the system once and consequences made by the attack exist for a long period. Cyber security system is needed in every single process associated with the organization [5].

The presented paper addresses the problem of resilient detection of threat and more processing time in the network. It also considers the major issue of system replay issues. The overtake of cyber threat dominates the system operation. The presented paper discusses the detailed Literature background in Section II. Followed by Section III, discuss the tool selection and impacts of specific tool to deep analysis of the proposed idea. Section IV. Discuss the design methodology of the proposed model, followed by Section V. Results and discussions are made.

II. Literature Survey

K. Thakur et al., [6] discussed various cyber security threats in the form of data utilized, work enforcement, and protocols used to protect the user information. Emails and virus scanning assistance providing significant results are discussed. He investigated the peer review of papers with similar ideologies and found out the need for password protection frameworks in cyber security systems. The paper provides the knowledge on basic techniques to safeguard the IOT devices.

H. Bennasar et al., [7] evaluated a journal addressing various network security algorithms utilized for the cloud security threats. They discussed on cloud network model and its infrastructure that can be optimized. The cloud services such as SaaS, PaaS, IaaS being discussed and various threats such as data loss, Account hacking, malwares in the network and device malfunction etc.

R. Das et al., [8] discussed a journal stating various machine learning algorithms that incorporated with the development of cyber security models. A detailed survey has been framed and the comparative performance on each algorithm in terms of accuracy and error rate is evaluated. OneR algorithm, Naïve Bayes, random forest algorithms are discussed together to find the better performing model. Artificial neural network models and clustering algorithms are reviewed.

I. Duic et al., [9] highlighted the international cyber security standards in their research work stating the cyberspace implications in the global scenario. The paper explained much about the security risk and challenges to be expected in fast growing internet world. They elaborated the implementation as two phases. In the first phase they discussed about the summarized list of attacks that target the IOT devices. The second phase of discussion is about the intrusions present in the network. Throughout the paper information related to malwares, denial of service, unauthorized access and many key impacted factors are discussed.

D. Ratasich et al., [10] evaluated a research work that discusses the state of art approach of various existing cyber security methodology on fault detection process, self-healing methodology and anomaly detection systems that is behind the resilience of cyber security devices. The amount of robustness is considered as one of the main attribute of sustaining the resilient issue sin cyber security systems.

M. Saharkhizan et al., [11] evaluated a deep recurrent neural network model for detecting the cyber-attacks in internet controlled devices using traffic data of network. Deep LSTM model is associated with the ensemble detectors, to get the effective outcome on malware detection and achieved a rate of 99% of accuracy. The developed model is also tested with the Modbus data of network traffic. The correlated patterns helpful for us to determine the factors related with the cyber-attacks.

A. Sivanathan et al., [12] developed a challenging network classification model that is pre-trained by historical data streams of IoT traffics. They developed a one stop customized behavioral model for detecting the network performance. Automatic noise filtering is done before evaluating the traffic path. IOT devices connected with the certain traffic are intended to provide the demonstration on their individual performance.

Na Liu et al., [13] evaluated new perceptions on cyber-attacks over connected vehicles and autonomous vehicles. The research aimed to provide information on lack of knowledge on analyzing the cyber-attacks in user authentications, licensing part of CAV (connected and autonomous vehicles). Safety, awareness, responsibility education and trust factors are considered.

Isabel Arend et al., [14] they reported a detailed analysis on malwares and its attacks, to the cybercrimes feasibilities happening because of the lack of user security. The study reveals active risk and passive risk on cyber security behaviors. They concurrently evaluate the difference between the two risks present in the cyber security systems. The research underwent various challenges towards the discussions on theoretical and practical implications.

A. Related Work

Wang et al., [15] Apart from firewall security, conventional security systems are required for Supervisory control and Data acquisition (SCADA) is discussed in which deep learning approach is evaluated. It detects the malicious attacks in the SCADA environment and investigates the protection criteria of cyber-attacks detection process. Alkahtani et al., [16] Botnet attacks are one of the serious problems in cyber systems, which threated the motion-less IoT devices. Empirical research was made on the malicious pattern detection using Convolution neural network (CNN) combined with Long-Short term memory Neural network (LSTM). It has the improved ability to detect the BOTNET attacks, with accuracy of 90.88%. F. Hossain et al., [17] a reliable cyber-attack detection model with ensemble classification model is evaluated. Gradient boost algorithm and random forest algorithm is converged for effective detection [18]. The static test and evaluation of cyber threat detection using LASSO classifier achieved the accuracy of 99% maximum [19].

III. System Design

The major addressable issue of Cyber security attacks in publicly available networks is the resilient response of the system for the injected issues, and those remains in the system for prolonged duration and permanently damages the organization credentials. The overall processing delay on attack detection and system replay issues are addressed here.

The novel cyber-attacks detection protocol is developed using the hybrid combination of cascaded R2CNN in which the regression analysis is cascaded in decision making with the recurrent convolution neural network model. The system design is implemented using MATLAB IDE by utilizing the regression tools and deep neural network toolbox. The system prediction model is developed in a recurrent fashion to test the weightage and apply the bias result again with the inputs that adapt the values.

IV. Design Methodology

A. Design Summary

Fig. 2 shows the architecture of R2CNN cascaded model. The preprocessing part of the system uses adaptive principle component analysis and Singular value decomposition model combined to process the raw data. The LYSIS dataset consists of eight columns of received data. The dataset holds the parameters of connected device in the IOT modems. The feature mapped variables are applied to gradient boosted regression model in which the unsupervised analysis is made. The closely related parameters are plotted in regression plots as shown in Fig. 4(a) (b) and (c). The prediction result is based on two decisions. The decision one is created using the regression method.

Followed by regression cascaded fashion of recurrent Convolution neural network (R2CNN) is implemented. The dataset patterns after the feature mappings form a feature vectors. These vectors are resized to constant matrix dimension and fetched to RCNN model. R2CNN model consists of Input layer of size [100x1], max pooling layer, convolution layer (1x10: Stride) and fully connected layer (384x5 layers). The data is pattern compared at each stages of recurrent neural network.

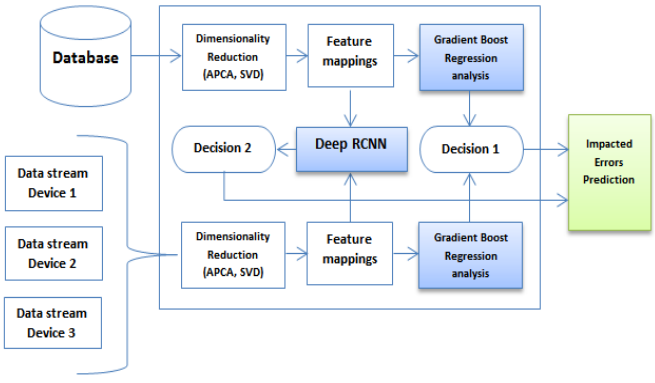


Fig. 2. Architecture Diagram of Cascaded R2CNN Prediction Model.

The loop of recurrence is varied for improving the performance. Every stage of analysis the weighted bias of the comparative result is updated. The profound R2CNN network is executed to distinguish the example coordinate between the data base and the test contribution of the IOT information. These examination results are additionally put away as large information stockpiling focuses for future analysis. The proposed framework utilizes four unique calculations and its similar outcomes.

The exactness of the forecast framework is determined utilizing the disarray network delivered toward the finish of R2CNN simulation. In case of threat pattern that does not correlate with any of the database learned patterns, then it is considered as the new threat and it is intended for learning purpose. Further new entries of cyber-attack patterns are also considered here. The disarray framework creates the boundaries, for example, Truepositive rate (TP), True negative rate (TN), false positive rate (FP) and False negative rate (FN). The exactness is determined utilizing the equation given in equation(1).

$$\text{Accuracy} = [TN+TP/(TN+TP+FN+FP)] \quad (1)$$

V. Result and Discussion

Fig. 3 shows the accuracy graph of deep R2CNN model that depicts the improved accuracy with respect to the increase in iteration of the analysis. The pre-trained data compared with the test input pattern from the device. In case of maximum correlation with the malware data in the database, the prediction system provides the label as output showing malware detection as message box showing the impacted factor such as prolonged duration, malicious application found or unauthorized notifications found, etc. as shown in Fig. 5.

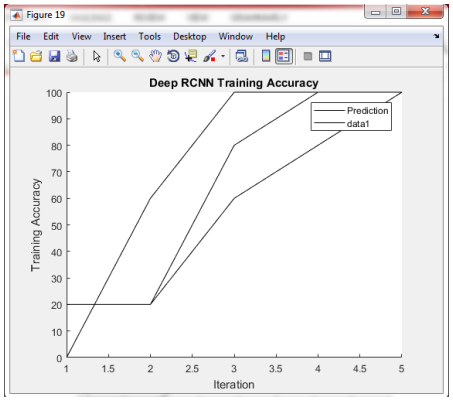


Fig. 3. Training Accuracy of Cascaded R2CNN.

Fig. 4 shows the regression analysis graph of direct method applied to the input data, 4(b) shows the regression analysis plot using Matlab inbuilt function, and 4(c) shows the regression analysis plot using gradient descent boosted regression method as proposed. The result of the regression model shows the maximum relative values towards the diagonal line. The regression result is considered as the decision 1 for final prediction. The detailed algorithm steps are shown.

Algorithm: Cascaded R2CNN

1 *Get input data*
2 *Extract Features X=features(data)*
3 *Compute Gradient Descent Regression(X);*
4 *Fetch to RCNN modeland store Result = weightage*
5 *Check bias (Result)& apply to Y*
6 *Repeat loop*
7 *Compare X and Y for max(Match score)*
8 *Call parameters(max(match_data))*
9 *Display parameters p1,p2,p3*
10 *Repeat all steps.*

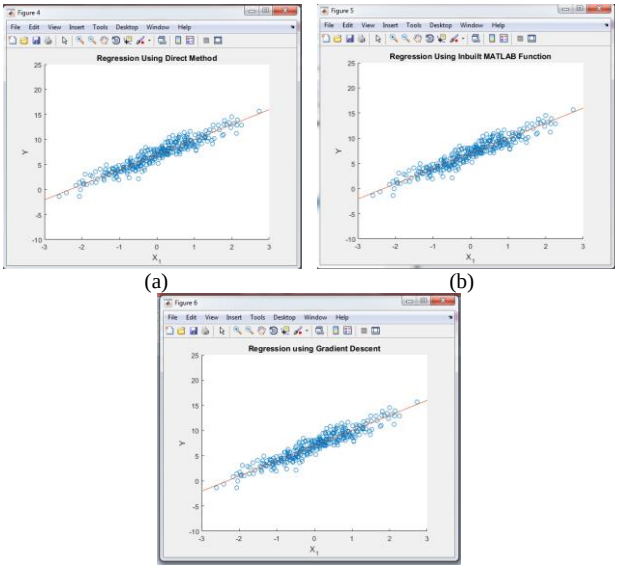


Fig. 4. (a) Regression Graph using Direct Method (b) Regression using Inbuilt Matlab Function (c) Regression using Gradient Descent Boosted Method.

Table I shows the Comparison results of Existing works and Proposed Cascaded R2CNN architecture. The author in [11] developed Cyber Threat detection in Traffic data using LSTM algorithm and achieved the accuracy of 99%. In [16], developed a LSTM with CNN configuration and achieved the accuracy of 90.88%. In [17], the author discusses the Massive cyber Protection system with Gradient Boost algorithms and Random forest models. Accuracy of 99% is achieved by the author. These studies helpful in deriving the proposed Cascaded architecture that combine the regression results as well as Convolution operation.

TABLE I. COMPARISON OF EXISTING WORK AND PROPOSED CR2CNN

S No	References	Concept	Algorithm	Accuracy
1	M. Saharkhizan et al., [11]	Cyber Threats detection in traffic data	LSTM	99%
2	Alkahtani et al., [16]	Cyber threat detectionin SCADA	LSTM-CNN	90.88%
3	F. Hossain et al., [17]	Massive cyber system protection	GBR-RF	99%
4	Proposed Work	Cyber threat detection in Smart city Dataset- LYSIS	CR2CNN	99.2%

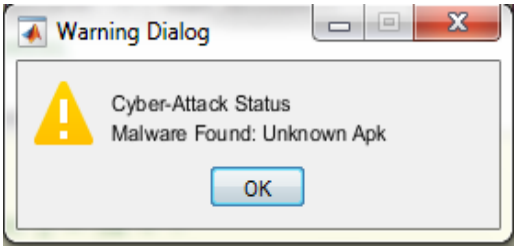


Fig. 5. Alert Notification.

Fig. 5 shows the Cyber-attack status after testing the input. The output also shows the parameter that impacted for malware. Example the above alert message displays the unknown APK found as parametric result that leads to cyber-attacks. The proposed algorithm focused on detecting the type of cyber-attack in short span of time. The resilience issue addressed is reduced here; comparatively the proposed system achieved the delay of 10.77 seconds.

Fig. 6 shows the mean square error of the three methods of regression analysis. From the above figure it is clear that gradient descent boosted model provides less error rate.

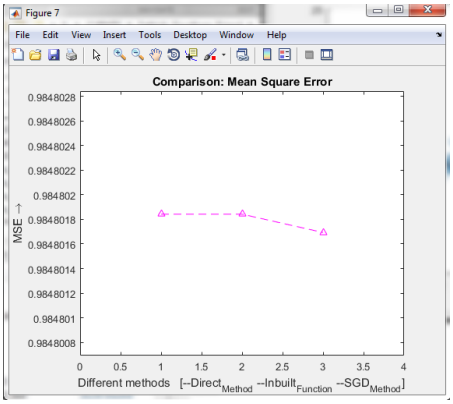


Fig. 6. Mean Square Error of Proposed Method.

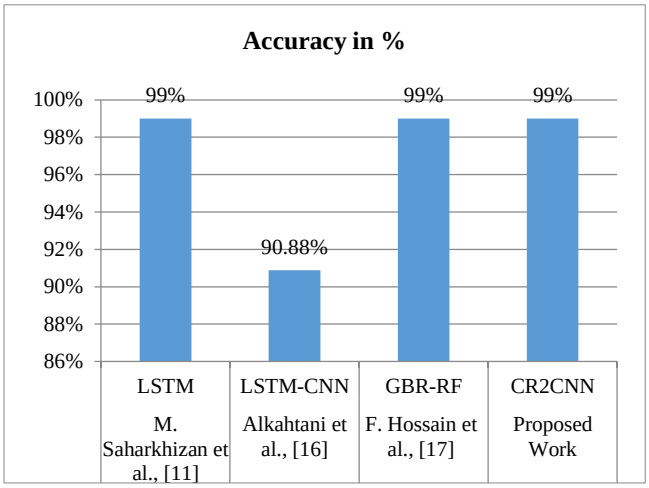


Fig. 7. Comparative Accuracy of Existing and Proposed Cyber-attack Detection Models.

Fig. 7 shows the comparison results of accuracy with various existing implementations. Considering the static dataset of LYSIS smart city, the presented system derives

maximum focus on reducing the processing delay for earliest prediction. The Cascaded R2CNN stands in the benefit of less computation steps. The stacked operation ensures the results with high confidence. Accuracy is high for the static approach anyway the present research need to be extended with dynamic approach on few real time threats in massive cyber physical systems. The proposed model tests the reliable communication in small enterprise level. Dynamic approach obviously extended for global discussion.

VI. Conclusion

Cyber-attack tracking and prevention is mandatory in almost every internet connected platforms. As an initiative the proposed system is focused on deriving a novel methodology to detect the most impacted leads for cyber-attacks in IoT networks. The proposed system consists of Novel algorithm named Cascaded R2CNN. Provided with LYSIS dataset, the developed model detects the Cyber threatening patterns. The addressed issue of early prediction is given priority here. The system achieved the early prediction of Cyber-attack patterns within 10.77 seconds. The proposed prediction model achieves 99.2% accuracy towards the given input dataset. Further the system modeled here is reliable for static testing of IoT networks in enterprise level. The proposed research work needs to be extended to achieve dynamic results by considering multiple environments and different patterns of Cyber-Threatening patterns.

References

[1] Sarker, I.H.; Abushark, Y.B.; Alsolami, F.; Khan, A.I. IntruDTree: AMachine Learning Based Cyber Security Intrusion Detection Model. Symmetry 2020, 12, 754.

[2] R. Kozik and MichałChoraś “,Machine Learning Techniques for Cyber Attacks Detection”, published Springer international year 2014.

[3] Aman, W. and Shukaili, J., 2021. A Classification of Essential Factors for the Development and Implementation of Cyber Security Strategy in Public Sector Organizations. International Journal of Advanced Computer Science and Applications, 12(8).

[4] Majeed, R., Abdullah, N. and Mushtaq, M., 2021. IoT-based Cyber-security of Drones using the Naïve Bayes Algorithm. International Journal of Advanced Computer Science and Applications, 12(7).

[5] LatifaAlzahrani, “Statistical Analysis of Cybersecurity Awareness Issues in Higher Education Institutes” International Journal of Advanced Computer Science and Applications(IJACSA), 12(11), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0121172.

[6] K. Thakur, M. Qiu, K. Gai and M. L. Ali, "An Investigation on Cyber Security Threats and Security Models," 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing, New York, NY, 2015, pp. 307-311, doi: 10.1109/CSCloud.2015.71.

[7] H. Bennasar, M. Essaaidi, A. Bendahmane and J. Ben-othman, "State-of-the-art of cloud computing cyber-security," 2015 Third World Conference on Complex Systems (WCCS), Marrakech, 2015, pp. 1-7, doi: 10.1109/ICoCS.2015.7483283.

[8] R. Das and T. H. Morris, "Machine Learning and Cyber Security," 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Kolkata, 2017, pp. 1-7, doi: 10.1109/ICCECE.2017.8526232.

[9] I. Duic, V. Cvrtila and T. Ivanjko, "International cyber security challenges," 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2017, pp. 1309-1313, doi: 10.23919/MIPRO.2017.7973625.

[10] D. Ratasic, F. Khalid, F. Geissler, R. Grosu, M. Shafique and E. Bartocci, "A Roadmap Toward the Resilient Internet of Things for Cyber-Physical Systems," in IEEE Access, vol. 7, pp. 13260-13283, 2019, doi: 10.1109/ACCESS.2019.2891969.

[11] M. Saharkhizan, A. Azmoodeh, A. Dehghantanha, K. -K. R. Choo and R. M. Parizi, "An Ensemble of Deep Recurrent Neural Networks for Detecting IoT Cyber Attacks Using Network Traffic," in IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8852-8859, Sept. 2020, doi: 10.1109/JIOT.2020.2996425.

[12] A. Sivanathan, H. H. Gharakheili and V. Sivaraman, "Detecting Behavioral Change of IoT Devices Using Clustering-Based Network Traffic Modeling," in IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7295-7309, Aug. 2020, doi: 10.1109/JIOT.2020.2984030.

[13] Na Liu, AlexandrosNikitas, Simon Parkinson,Exploring expert perceptions about the cyber security and privacy of Connected and Autonomous Vehicles: A thematic analysis approach,Transportation Research Part F: Traffic Psychology and Behaviour, Volume 75,2020,Pages 66-86,ISSN 1369-8478,https://doi.org/10.1016/j.trf.2020.09.019.

[14] Isabel Arend, AsafShabtai, TaliIdan, RutyKeinan, YoellaBereby-Meyer, Passive- and not active-risk tendencies predict cyber security behavior, Computers & Security, Volume 97, 2020, 101964, ISSN 0167-4048.

[15] Wang, W., Harrou, F., Bouyeddou, B. et al. A stacked deep learning approach to cyber-attacks detection in industrial systems: application to power system and gas pipeline systems. Cluster Comput 25, 561–578 (2022). https://doi.org/10.1007/s10586-021-03426-w.

[16] Alkahtani, H. and Aldhyani, T., 2021. Botnet Attack Detection by Using CNN-LSTM Model for Internet of Things Applications.Security and Communication Networks, 2021, pp.1-23.

[17] F. Hossain, M. Akter and M. N. Uddin, "Cyber Attack Detection Model (CADM) Based on Machine Learning Approach," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 567-572, doi: 10.1109/ICREST51555.2021.9331094.

[18] K. R. Choo, M. Conti and A. Dehghantanha, "Special Issue on Big Data Applications in Cyber Security and Threat Intelligence – Part 1," in IEEE Transactions on Big Data, vol. 5, no. 3, pp. 279-281, 1 Sept. 2019, doi: 10.1109/TBDATA.2019.2933039.

[19] S. Merat and W. Almuhtadi, "Artificial intelligence application for improving cyber-security acquirement," 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), Halifax, NS, 2015, pp. 1445-1450, doi: 10.1109/CCECE.2015.7129493.

A Secure and Robust Architecture based on Mobile Healthcare Applications for Patient Monitoring Environments

Shaik Shakeel Ahamad, Majed Alowaidi
Department of Information Technology, College of Computer and Information Sciences
Majmaah University, Al-Majmaah, 11952, Saudi Arabia

***Abstract*—The recent outbreak of COVID-19 pandemic realized the importance of patient monitoring environments, Mobile Healthcare Applications (MHA) plays very crucial role in the successful implementation of patient monitoring environments. Existing MHA's in the realm of patient monitoring environments are prone to repackaging attacks; do not ensure security, application security and communication security. This paper proposes a secure and robust architecture for mobile healthcare applications in patient monitoring environments ensuring end to end security ensuring all the security properties by overcoming repackaging attacks which are very vital for success of mobile healthcare applications. We implemented our proposed protocol in Android Studio, Kotlin is designed to interoperate fully with Java. ECDH Key exchange algorithm is used for key exchange between MHA in patient's smart phone and MHA in the hospital TPM. We created an EC key pairs (NIST P-256 aka secp256r1) at patient's MHA and MHA of hospital TPM by using ECDH and we created a shared AES secret key. AES with GCM mode used for encryption and decryption of patient data.**

***Keywords*—Mobile healthcare applications (MHA); UICC (universal integrated circuit card); Kotlin language; android studio; ECDSA (elliptic curve digital signature algorithm); GCM mode; end to end security**

I. Introduction

The speedy development of information and communication technology (ICT) infrastructures are playing very important role in offering innumerous opportunities for efficient and affordable mobile health solutions. Mobile health solutions help in delivering healthcare anywhere and at any time overcoming geographical barriers, these services are a boon for the patients living in remote areas where health care facilities are not accessible. Mobile Healthcare Applications (MHAs) plays vital role in the successful implementation of mobile health solutions. Many Mobile Healthcare Applications (MHAs) are available in the market helping hospitals, doctors and patients. MHAs are available to assist hospitals, doctors in managing and in monitoring patients and in making clinical decisions. Smartphones and MHAs provide the following benefits to all the stakeholders especially patients, doctors and hospital staff as they ensure accuracy and efficiency. The author in [1] systematically assessed the consequences of cyber threats on health care. The security of user privacy information is very important for system deployment and operation [2]. The authentication process of Telecare Medical Information Systems (TMIS) occurs in a public channel, which is prone to attacks. Attackers can disrupt the authentication process through eavesdropping, interception, and forgery method, and launch malicious attacks such as forgery attacks, replay attacks, and side-channel attacks. These attacks can lead to malicious access and loss of data. Future MHAs are expected to include larger databases helping in making clinical decisions. COVID-19 pandemic realized the importance of patient monitoring environments, Mobile Healthcare Applications (MHA) plays very crucial role in the successful implementation of patient monitoring environments. During COVID-19 pandemic health information system became the primary target of cybersecurity attacks [3]. The health care industry should be prepared to overcome cyberattacks. The system can be protected from attacks by designing a secure identity authentication scheme and intrusion detection technology [4]. Among the main concerns in health monitoring frameworks are: reliability in making clinical decisions and security and privacy of data. Existing MHA's in the realm of patient monitoring environments do not ensure application security and communication security. Existing mobile healthcare monitoring solutions does not ensure Application security and communication security, Patient's privacy, not compliant with HIPAA standard and prone to repackaging attacks. This article's organization is as follows: In Section II discusses related work in the realm of secure mobile healthcare. Section III proposes a Secure Mobile Healthcare framework. Section IV presents an experimental setup and results, and Section V compares our proposed work with the related works. Section VI provides discussion of the proposed framework, and Section VII concludes the paper.

II. Related Work

[5] monitors blood pressure, with a unique look and feel for monitoring heart health, which communicates with Bluetooth, so it is easy to share and store patient's records. The Omron HeartAdvisor mobile app [6] allows to transfers blood pressure readings smartphone based healthcare application. But both the solutions have the following limitations

a) End to End security is not ensured.

b) Patient's privacy is not ensured.

c) Not compliant with HIPAA standard.

d) Does not ensure application security

e) Does not ensure Communication security

f) These solutions are vulnerable to repackaging attacks

The author in [7] proposes an authentication scheme in Telecare Medical Information System (TMIS) based on Physical Unclonable Function (PUF) and Elliptic Curve Cryptography (ECC) technology. But this solution has no clarity:

a) How the ECC technology can encrypt the messages in the real time.

b) How the healthcare application overcomes reverse engineering attacks?

The author in [8] proposes healthcare systems with mutual authentication protocol thereby ensuring location privacy with low computation and storage costs, but this work also does not ensure application security and communication security. The author in [9] proposes a Cloud-IoT based healthcare system that uses a lightweight user authentication scheme, but this work do not ensure end to end security and prone to repackaging attacks. According to market watch, the Application Security Market will cross US$ 11 billion by 2024 globally [10]. According to marketsandmarkets IoT medical devices are will reach USD 63.43 billion by 2023 globally [11]. IoT medical devices are being used by many patients all around the globe as they make the life of patients easy and is evident from the predictions from marketsandmarkets [12], but these devices should be made secure right from the manufacturing phase of these devices which is the responsibility of the manufacturer. IoT medical devices use healthcare applications and applications need to be portable and secure, the security of these applications is the responsibility of the hospitals and the government. Healthcare data is kept in the hospital database and it is the responsibility of the hospitals and the government to keep the data secure thereby ensuring HIPAA regulations. In order to be HIPAA complaint network security should be ensured, i.e. protecting data at rest and during transit. This is the core motivation for this work. PhysioDroid [13] is an advanced system for remote monitoring of patient's health. The PhysioDroid system has the following:

1) A monitoring device transmits the collected readings.

2) A smartphone, data collector application for medical diagnosis and for health alerts.

3) Stores data from multiple sources.

The author in [14] discusses transport issues in the mobile Healthcare applications, proposes a platform for testing and finally proposes solutions to overcome these attacks. The author in [15] discusses server side security concerns and vulnerabilities in the mHealth apps and compares with the applications in other realms. The author in [16] proposes a data encryption solution for mobile health apps (DE4MHA). Following are the limitations of the existing research works in the realm of Mobile Healthcare Applications (MHA):

a) Application security and communication security is not ensured

b) Data in the hospital is not secure.

c) Patient's privacy is not ensured.

d) Not compliant with HIPAA standard.

e) Does not ensure application security

f) Does not ensure Communication security

g) Existing MHA's are vulnerable to repackaging attacks

The author in [17] proposes a new self-defending code (SDC) approach which encrypts parts of the app code at compile time and dynamically decrypts the ciphertext code at run-time but this work does not ensure the security of keys.

Following are the contributions made by our research work:

a) We have proposed a secure architecture from the UICC (Universal Integrated Circuit Card) of the patient's smart phone and hospital server and a secure protocol is proposed in the realm of Patient Management and Monitoring.

b) In our proposed healthcare framework MHAs overcomes repackaging attacks code obfuscation, code attestation and by enabling self-signing restrictions.

c) We have proposed a secure healthcare protocol ensuring all the security properties.

d) Compared our proposed healthcare system with the existing real time Mobile Healthcare Application solutions and existing research works in mobile healthcare and found to be better than these solutions and

e) We successfully implemented our proposed protocol in Android Studio and found to be better than the existing solutions.

f) Proposed healthcare framework overcomes known attacks.

III. Proposed Healthcare Framework

In order to overcome the existing MHA and research works in the realm of Mobile Healthcare we propose a secure interaction between the MHA in the UICC of the patient's smartphone and the TPM of the hospital. Patient (P), Doctor (D), Hospital (H), Sensor (S), MHA in sensor, UICC in Smartphone and MHA in the UICC are the entities involved in the proposed framework. Existing MHAs are installed in the smart phone which can be compromised by malware, so we propose our secure framework in the SE of the patient's smartphone referred as UICC. Sensor (S) contains a SE, SE contains MHA collecting health information. This Sensor (S) MHA shares a symmetric key with the MHA in UICC of the patient's smartphone as shown in [12] and MHA of patient's smartphone shares a symmetric key with the MHA of TPM at hospital. UICC and MHAs in the UICC of the patient's smartphone are personalized by the TPM at hospital as shown in [18] Over-The-Air (OTA). TPM of the hospital is personalized by the hospital. Fig. 1 shows the interaction between the Patient's MHA (which is in the UICC of smartphone) and MHA of Hospital (which is in the Hospital TPM). There are nine layers at the both sides i.e. MHA, HTTPS (HTTPS Request and HTTPS Response), TLS, TCP, IP, BIP, SCP 102 223, SCP 102 221 and ISO 7816-3/4. MHA of patient's smartphone encrypts the messages with a symmetric key shared between MHAs of patient's smartphone

and TPM at hospital. HTTPS encrypts all the messages exchanged between patient's smartphone and TPM at hospital. Communication security is ensured using TLS a secure tunnel is established between patients.

UICC of smartphone and Hospital TPM, TCP ensures end to end reliability, IP is a protocol used at the network layer and BIP is a mechanism at the interface between the UICC and the smartphone providing access to the data bearers supported by the smartphone. ISO and the IEC jointly manages ISO/IEC 7816 standard. By using our proposed secure architecture end to end security and reliability is ensured in the information exchange between the patient and the hospital. Table I shows the notations used in the paper. Fig. 2 shows the steps involved in patient monitoring protocol.

Step 1: Sensor (S) collects patient's readings and sends it to the UICC of the patient's smartphone at regular intervals via Bluetooth Low Energy (BLE); in order to overcome BLE vulnerabilities, MHA in Sensor (S) encrypts the data sent to the MHA in the UICC of the smartphone (patient (P)). Patient's readings are encrypted with the shared symmetric key between the MHA of the Sensor (S) and MHA in UICC (P). Our proposed framework overcomes BLE vulnerabilities as our MHA's code is obfuscated by the MHA manufacturer and attested by the Certifying Authority (CA) and imposes self-signing restrictions, in addition to these data transmitted from the sensor (S) is encrypted using the symmetric key shared between sensor's MHA and the MHA of the patient (P). Data encryption prevents MITM and eavesdropping attacks. A secure link is established between the sensor's MHA and MHA in the UICC of the patient ensuring application security (symmetric key) and communication security (using SSL/TLS).

$$S \rightarrow P: \{PD, T_S, N_S, ID_P, ID_S, LOC_P\}SK_{PS}$$

Step 2: UICC (P) forwards the received message to the hospital's Trusted Platform Module (TPM) after decrypting the received message.

$$P \rightarrow H: \{PD, LOC_P, T_P, N_P, ID_P, ID_H\}SK_{PH}$$

Step 3: If the readings are abnormal then "H" shares patient's location to the ambulance

$$H \rightarrow D: \{PD, ID_P, LOC_P, T_H, N_H, ID_H\}SK_{HD}$$

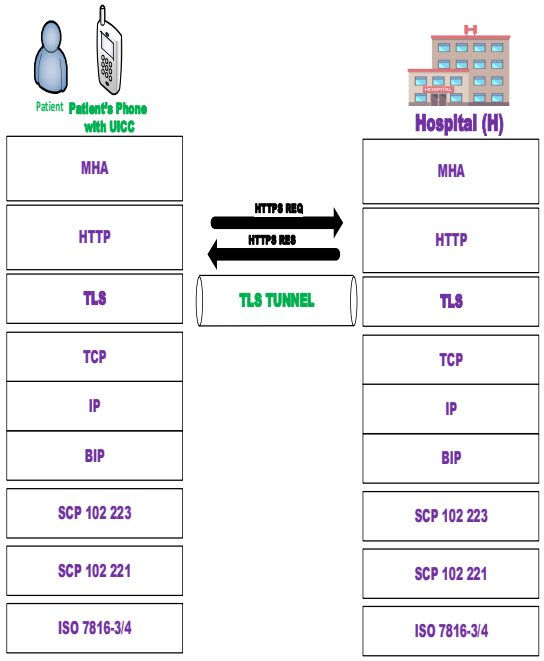


Fig. 1. Interaction between the Patient and Hospital.

TABLE I. NOTATIONS

Notation	Full Form/Meaning	Notation	Full Form/Meaning	Notation	Full Form/Meaning
MHA	Mobile Healthcare Application	**OTA**	Over The Air	**ID_P**	Identity of Patient
HTTPS REQ	Hypertext Transfer Protocol Secure Request	**AES**	Advanced Encryption Standard,	**PD**	Patient Data
HTTPS RES	Hypertext Transfer Protocol Secure Response	**ECDH**	Elliptic-curve Diffie–Hellman	**P**	Patient
TLS	Transport Layer Security	**ECDSA**	Elliptic Curve Digital Signature Algorithm	**ID_H**	Identity of Hospital
IP	Internet Protocol	**SE**	Secure Element	**ACK**	Acknowledgment
BIP	Bearer Independent Protocol	**UICC**	Universal Integrated Circuit Card	**SK_{PH}**	Symmetric Key shared between Patient and Hospital
SCP 102 223	Smart Card Platform 102 223	**GCM**	Galois/Counter Mode	**SK_{HD}**	Symmetric Key shared between Hospital & Doctor
SCP 102 221	Smart Card Platform 102 221	**NIST**	National Institute of Standards and Technology	**SK_{PS}**	Symmetric Key Shared between Patient & Sensor
ISO 7816-3/4	International Organization for Standardization (ISO)	**EC Key Pair**	Elliptic Curve Key Pair	**T_P**	Timestamp generated by Patient
H	Hospital	**HIPAA**	Health Insurance Portability and Accountability Act	**T_H**	Timestamp generated by Hospital
P	Sensor	**IoT**	Internet of Things	**T_S**	Timestamp generated by Sensor
D	Doctor	**IV**	Initialization Vector	**N_P**	Nonce generated by Patient
N_S	Nonce generated by Sensor	**N_H**	Nonce generated by Hospital	**LOC_P**	Location of the Patient

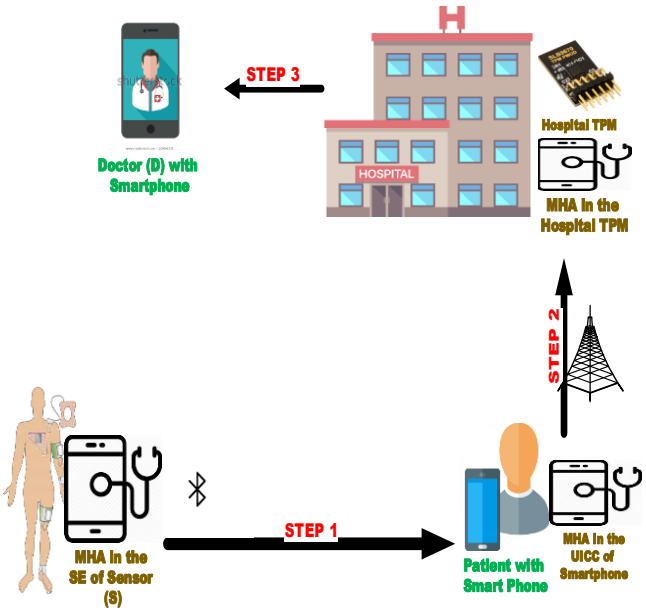


Fig. 2. Proposed Healthcare Protocol.

IV. Experimental Setup and Results

We implemented our proposed protocol in Android Studio using Kotlin language; it was designed to interoperate fully with Java. ECDH Key exchange algorithm is used for key exchange between MHA in patient's smart phone and MHA in the hospital TPM. ECDSA, digest algorithm used is SHA-256 and AES symmetric encryption algorithm are used to ensure all the security properties. We created an EC key pairs (NIST P-256 aka secp256r1) at patient's MHA and MHA of hospital TPM by using ECDH and we created a shared AES secret key. AES with GCM mode used for encryption and decryption of patient data, Fig. 3 and 4.

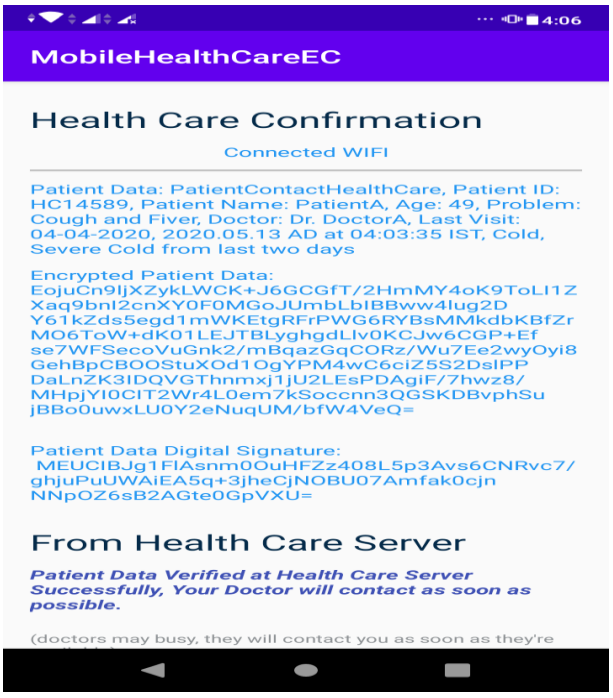


Fig. 3. Encrypted Patient's Data Transferred to Hospital.

Fig. 4. Confirmation of Patient Data at Hospital.

V. Comparison with Related Work

We have proposed a secure architecture from the UICC of the patient's smart phone and hospital server and a secure protocol is proposed in the realm of Patient Management and Monitoring. Proposed healthcare framework MHAs overcomes repackaging attacks code obfuscation, code attestation and by enabling self-signing restrictions. In addition to these proposed secure healthcare protocol ensuring all the security properties. Finally, we have carried out our experiments in Android Studio and found to be better than the existing solutions by overcoming all the known attacks. This section highlights the comparative analysis of the proposed system with the eight existing real time MHA solutions and existing research works. Table II compares our proposed framework with the existing research works in this realm with the following features and found to be better than the existing research works.

a) Confidentiality: Confidentiality is ensured using a symmetric key at the application level which is shared between the entities.

b) Authentication: MHAs are authenticated using their respective certificates and moreover symmetric keys are shared among the entities involved in the framework.

c) Overcomes Tampering of Messages: Messages are encrypted using the shared symmetric key and data is hashed thereby ensuring the integrity of the exchanged messages.

d) Compliant to HIPAA Standard: Messages are encrypted using the shared symmetric key and communication security is also ensured using SSL/TLS protocol

e) Application Security: MHAs in both the sensor and UICC (of the patient) are protected by password. In addition to this MHAs are protected from repackaging attacks by implementing code obfuscation, code attestation and by enabling self-signing restrictions.

TABLE II. COMPARATIVE ANALYSIS WITH RELATED WORK

Research Works Features	**[5]**	**[6]**	**[13]**	**[14]**	**[15]**	**[16]**	**Our Proposed**
Confidentiality	No	No	No	Yes	Yes	Yes	Yes
Authentication				Yes	Yes	Yes	Yes
Overcomes Tampering of messages	No	No	No	Yes	Yes	Yes	Yes
Compliant to HIPAA standard	No	No	No	No	No	No	Yes
Ensures Application Security	No	No	No	No	No	No	Yes
Ensures Communication Security	No	No	No	Yes	Yes	No	Yes
Overcomes Heartbleed Vulnerability	No	No	No	Yes	Yes	No	Yes
Overcomes BLE vulnerabilities	No	No	No	No	No	No	Yes
Overcomes Replay Attacks	No	No	No	No	No	Yes	Yes
Overcomes Man-In-The-Middle Attacks	No	No	No	No	No	No	Yes
Overcomes Impersonation Attacks	No	No	No	No	No	No	Yes
Overcomes reverse engineering attacks	No	No	No	No	No	No	Yes

f) Communication Security: TLS protocol at the communication layer ensures communication security.

g) Overcomes heartbleed vulnerability: Our proposed system uses newer versions of TLS certificates signed by the Certifying Authority (CA). Patients' private keys are secure which is vital to overcome this vulnerability. So our proposed system overcomes Heartbleed vulnerability.

h) Overcomes BLE Vulnerabilities: Sensor (S) and the UICC of the smartphone communicates at regular intervals through Bluetooth Low Energy (BLE). Our proposed framework overcomes BLE vulnerabilities as our MHA's code is obfuscated by the MHA manufacturer and attested by the Certifying Authority (CA) and imposes self-signing restrictions. So our proposed system overcomes BLE vulnerabilities.

i) Overcomes Replay Attacks: Encrypted messages containing timestamps and nonce helps in s overcoming replay attacks

j) Overcomes Man-in-The Middle Attacks: Encrypted messages containing timestamps and nonce helps in s overcoming MITM attacks.

k) Overcomes Impersonation Attacks: Our proposed system overcomes an impersonation attack as the attacker will be unsuccessful in generating session keys.

l) Overcomes Reverse Engineering Attacks: MHAs are protected from repackaging attacks by implementing code obfuscation, code attestation and by enabling self-signing restrictions on MHAs.

VI. DISCUSSION

Health care industry is the main target of attackers as the existing healthcare solutions are very vulnerable. MHAs are updated through unreliable sources, so the security of these solutions is compromised putting patent's data in risk. So MHAs should be personalized and updated by the hospital after authenticating each other. Following are the recommendations for secure patient monitoring environments:

a) Healthcare solutions should ensure all the security properties.

b) Healthcare solutions should be compliant to HIPAA standard.

c) MHAs should withstand reverse engineering attacks

d) MHAs should encrypt the patient's data/information

e) Healthcare solutions should overcome all the known attacks.

f) Healthcare solutions should overcome BLE and heart-bleed vulnerabilities.

VII. CONCLUSION

Health care industry became the primary target of attackers during COVID-19 pandemic, so health care industry should overcome all the cybersecurity attacks. This paper proposes a secure and robust architecture for mobile healthcare framework in patient monitoring environment which is compliant to HIPAA standard, ensures all the security properties. Mobile Healthcare Applications (MHA) in our proposed healthcare framework overcomes reverse-engineering attacks. We implemented our proposed protocol in Android Studio, Kotlin using Kotlin language. ECDH Key exchange algorithm is used for key exchange between MHA in patient's smart phone and MHA in the hospital TPM. ECDSA, digest algorithm used is SHA-256 and AES symmetric encryption algorithm are used to ensure all the security properties. We created an EC key pairs (NIST P-256 aka secp256r1) at patient's MHA and MHA of hospital TPM by using ECDH and we created a shared AES secret key. AES with GCM mode used for encryption and decryption of patient data. Our proposed mobile healthcare framework overcomes all the known attacks.

ACKNOWLEDGMENT

The authors gratefully acknowledge the editor and the reviewers' helpful comments and suggestions, which have improved the presentation.

Funding: Dr. Shaik Shakeel Ahamad would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project No.R-2022-19.

Competing interests: The authors have declared that no competing interests exist.

REFERENCES

[1] D. W. Kim, J.Y. Choi and K.H. Han, "Risk management-based security evaluation model for telemedicine systems," *BMC Medical Informatics Decision Making*, vol.20, no.1, pp.1-14,2020. [doi: 10.1186/s12911-020-01145-7] [Medline: 32522216].

[2] S. Zhang, T. Yao, V.K.A. Sandor, T.H. Weng, W. Liang *et al.,* "A novel block chain-based privacy-preserving framework for online social networks," *Connection Science*, vol.33, no.3, pp.555-575, 2020 [doi: https://doi.org/10.1080/09540091.2020.1854181].

[3] C.M. Williams, R. Chaturvedi and K. Chakravarthy, "Cybersecurity Risks in a Pandemic," *Journal Medical Internet Research*, vol.22, no.9: e23692,2020 [FREE Full text] [doi: 10.2196/23692] [Medline: 32897869].

[4] W. Liang, L. Xiao, K. Zhang, M. Tang, D. He *et al.,* "Data fusion approach for collaborative anomaly intrusion detection in block chain-based systems," *IEEE Internet of Things Journal*, pp 1-1, 2021 [doi: https://doi.org/10.1109/JIOT.2021.3053842].

[5] Connected Omron Healthcare Products and Apps, 2021. [Online]. Available: https://omronhealthcare.com/service-and-support/connected-health/ [accessed 2021-07-02].

[6] HeartAdvisor, 2021. [Online]. Available: https://omronhealthcare.com/service-and-support/faq/omron-heartadvisor/ [accessed 2021-07-02].

[7] L. Xiao, S. Xie, D. Han, W. Liang, J. Guo *et al.,* "A lightweight authentication scheme for telecare medical information system," *Connection Science,* vol.33, no.3, pp.769-785, 2021 [doi: 10.1080/09540091.2021.1889976].

[8] A. Tewari and B.B. Gupta, "An internet-of-things-based security scheme for healthcare environment for robust location privacy," *International Journal of Computational Science and Engineering*, vol.21, no.2, pp.298-303, 2021 [doi:10.1504/IJCSE.2020.105742].

[9] G. Sharma and S. Kalra, "A Lightweight User Authentication Scheme for Cloud-IoT Based Healthcare Services," *Iranian Journal of Science and Technology Transactions of Electrical Engineering,* vol.43, pp.619-636, 2019. [doi:10.1007/s40998-018-0146-5].

[10] Application Security Market Size will reach US$ 11 Billion by 2024 - MarketWatch. [Online]. Available: https://www.marketwatch.com/press-release/application-security-market-size-will-reach-us-11-billion-by-2024-2019-05-06. [Accessed: 18-June-2021].

[11] S. Singh. IoT Medical Devices Market worth $63.43 Billion by 2023. [Online]. Available: https://www.marketsandmarkets.com/PressReleases/iot-medical-device.asp. [Accessed: 18-June-2021].

[12] S. Kungpisdan, B. Srinivasan and P.D. Le, "Lightweight Mobile Credit-Card Payment Protocol," in *Proc. 2003 INDOCRYPT*, New Delhi, India, pp. 295–308, 2003 [doi: https://doi.org/10.1007/978-3-540-24582-7_22].

[13] O. Banos, C. Villalonga, M. Damas, P. Gloesekoetter, H. Pomares *et al.,* "PhysioDroid: Combining Wearable Health Sensors and Mobile Devices for a Ubiquitous, Continuous, and Personal Monitoring," *The Scientific World Journal*. Volume 2014, Article ID 490824, 11 pages [doi: http://dx.doi.org/10.1155/2014/490824].

[14] J. Müthing, T. Jäschke and C. M. Friedrich, "Client-Focused Security Assessment of mHealth Apps and Recommended Practices to Prevent or Mitigate Transport Security Issues," *JMIR Mhealth Uhealth*, vol.5, no.10: e147, 2017 [doi: 10.2196/mhealth.7791].

[15] J. Müthing, R. Brüngel and C.M. Friedrich, "Server-Focused Security Assessment of Mobile Health Apps for Popular Mobile Platforms," *Journal Medical Internet Research,* vol.21, no.1*:* e9818, 2019 [doi: 10.2196/jmir.9818].

[16] B.M. Silva, J.JPC. Rodrigues, F. Canelo, I.C. Lopes and L. Zhou, "A Data Encryption Solution for Mobile Health Apps in Cooperation Environments," *Journal Medical Internet Research,* vol.15, no.4*:* e66, 2013 [doi: 10.2196/jmir.2498].

[17] K. Chen, Y. Zhang and P. Liu, "Leveraging Information Asymmetry to Transform Android Apps into Self-Defending Code Against Repackaging Attacks," *IEEE Transactions on Mobile Computing,* vol.17, no.8, pp. 1879-1893, 2018 [doi: 10.1109/TMC.2017.2782249].

[18] S.S. Ahamad, V.N. Sastry and S.K. Udgata, "Secure mobile payment framework based on UICC with formal verification," *International Journal of Computational Science and Engineering*, vol.9, no.4, pp. 355-370, 2014 [doi: 10.1504/IJCSE.2014.060718].

A Novel Predictive Scheme for Confirming State of Bipolar Disorder using Recurrent Decision Tree

Yashaswini K.A, Dr. Aditya Kishore Saxena
Dept of Computer Science and Engineering
School of Engineering, Presidency University
Bangalore, India

***Abstract*—Bipolar disorder is one of the most challenging illnesses where medical science is still struggling to achieve its landmark therapies. After reviewing existing prediction-based approaches towards investigating bipolar disorder, it is noted that existing approaches are more or less symptomatic and relates depression as sadness. It implies various theories that don't consider many precise indicators of confirming bipolar disorder. Therefore, this manuscript presents a novel framework capable of treating the dataset of depression and fine-tune it appropriately to subject it further to a machine learning-based predictive scheme. The proposed system subjects its dataset for a series of data cleaning operations followed by data preprocessing using a standard scale of rating bipolar level. Further usage of feature engineering and correlation analysis renders more contextual inference towards its statistical score. The proposed system also introduces a Recurrent Decision Tree that further contributes towards the predictive outcome of bipolar disorder. The outcome obtained showcases that the proposed scheme performs better than the conventional decision tree.**

***Keywords*—Bipolar disorder; depression; recurrent neural network; decision tree; prediction; sadness**

I. INTRODUCTION

The proposed paper is about performing prediction towards confirming if the subject is suffering from bipolar disorder on the basis of analysis of their mood-based statistics. Mood disorders are disturbing mental disorders that place an enormous burden on individuals, health care systems, and economies. The two most common mood swing disorders are depression and bipolar disorder (BD) [1-2]. Every year approximately 15% of the global population experiences depression, and 4% experiences bipolar disorder during their lifetime. An individual suffering from depression has marked symptoms of sadness, feelings of emptiness, anxiety, sleep disturbances, and a general lack of motivation and interest in activities [3]. Additional symptoms may include feelings of guilt or worthlessness, decreased energy, difficulty concentrating, suicide, and psychotic behavior. The severity of depression depends on multiple factors such as type of symptoms, duration, and its impact on the individual's social and occupational life. An additional serious mental illness, bipolar disorder, is also associated with depression [4]. An important difference between unipolar depression and bipolar disorder is that episodes of mania characterize the latter. This is characterized by inflated self-esteem, impulsive behavior, increased activity, decreased sleep, and goal-directed behavior [5]. Both conditions are inherited and are linked to genetic susceptibility to environmental elements, causing disruptions to internal biological and emotional states. BD symptoms are also associated with physical health problems, side effects, and social factors, and alcohol and drug abuse, which may also contribute to BD symptoms in all individuals [6]. Therefore, it is evident that increasing stress and unhealthy routine in our modern lifestyle result in BD, mood disorders, depression, and abnormalities. The proposed study focuses on predicting the status of BD in the next frame. However, it is quite very challenging to diagnose BD types in their early stages due to recurrent depressive episodes. With the advent of modern technologies such as artificial intelligence, machine learning, and other data-driven approaches, many research studies have explored the issue of BD and its countermeasure [7-8]. However, the analysis of the data streams generated from psychopathology poses a huge challenge due to its complex nature of episodic and temporal characteristics [9]. The studies have shown that data quality has a significant impact on learning or predictive models. This data type requires effective data modeling from the feature engineering viewpoint that exploits distinctive features [10]. Another problem is the selection of suitable data-driven approaches and learning models. Since there are various machine learning (ML) models, each has its own advantages and disadvantages. The selection of a suitable model depends on the dataset characteristics, problem context, model parameters, and configuration. The existing approaches have not considered this factor and adopted a stock ML function lacking additional features and modification.

Therefore, the proposed study address this problem using a unique mechanism of machine learning. The objectives defined for this purpose are as follows: i) to introduces an effective data modeling for feature analysis on bipolar disorder, ii) to present a novel neural network model for predicting the accurate state of bipolar disorder in patients, iii) to determine a suitable feature descriptors based on subjective observations and the clinical rating scales, and iv) to use the feature obtained for bipolar disorder prediction with respect to episodic and temporal, and v) to deploy a recurrent neural decision tree network for processing both episodic and temporal data. The remaining section of this paper is organized as follows: Section II presents the review of literature carried out in the context of bipolar disorder detection and prediction; Section III highlights the research gap identified based on the literature review; Section IV introduces the proposed system and

methodology adopted; Section V discusses the implementation procedure; Section VI presents analysis of the proposed model while Section VII presents result discussion and performance analysis. Finally, overall work is concluded in Section VIII.

II. Related Work

This section provides a brief analysis and review of existing research work in the context of bipolar disorder (BD) prediction in patients. The study also highlights the significant research gaps based on the review analysis.

Millions of people worldwide suffer from BD, which raises serious life-threatening concerns. Many data-driven approaches were introduced in the context of early diagnosis to provide timely treatment. The work carried out by Ceccarelli, and Mahmoud [11] has suggested a multimodal scheme to identify mental disorders from multimedia feeds. This scheme employs a recurrent neural network (RNN) developed based on a fusion technique that considers temporal information in the training phase. The study outcome claims to offer better predictive outcomes than the existing approaches, emphasizing precise modeling of temporal features. In the work of Morla et al. [12], optical coherence tomography (OCT) data is analyzed with different machine learning approaches (ANN) to benefit the diagnosis of BD in the early stage. The BD patients are identified based on the predive outcome of retinal thinning and thickness of the particular area. However, the study needs an extensive analysis to prove and validate the effectiveness of the presented scheme.

Apart from using OCT data, electroencephalogram (EEG) also provides crucial features to enhance the clinical diagnosis of bipolar disorder. In the study of Sotos et al. [13], the authors have collected EEG data, an extreme applied gradient boosting (XGB) technique to identify BD patients. The authors have also implemented other machine learning classifiers such as k-nearest neighbors (KNN), decision tree (DT), Naïve Bayes (NB), and support vector machine (SVM) for the comparative assessment. The study claims that XGB outperforms other machine learning classifiers regarding the accuracy, recall, and precision. However, the presented model is not scalable; it does not perform well on sparse data samples and is often prone to outliers. Chen et al. [14] also tried to explore the effectiveness of artificial intelligence with the neuroimaging modalities. The authors have used magnetic resonance images (MRI) to differentiate schizophrenia patients in this study. A discriminative analysis considers coarse-to-fine feature selection to extract the differences between different samples. In this model, SVM is implemented as a predictive model. Sun et al. [15] have utilized Single Nucleotide Polymorphism to obtain features associated with genetic markers. Further, these markers are combined with Convolutional Neural Network (CNN) for recognizing BP patients. The outcome shows that the model has achieved only a 79% percentage recognition rate and claimed to offer excellent performance.

The deep learning approach is also used in Li et al. [16] for automated diagnosis of BD, psychotic symptoms, and healthy controls. The authors have considered MRI data and constructed CNN to classify the gray matter density images. The outcome shows that this study has achieved an accuracy of 99% and provides the clinical basis for disease diagnosis and treatment. Though the model can extract features automatically, it requires precise modeling and selection of the layers. Also, the presented model suffers from huge computational complexity and is prone to overfitting and exploiting gradient. Similarly, a more complex learning model can be seen in Huang et al. [17], which is based on the joint approach of CNN and RNN to identify short-term mood disorders. In this work, speech responses were fed to CNN, which provides an emotion profile. Further, RNN is applied to emotion profiles to determine the temporal relationship in emotion profiles to detect mood swings. The work carried out Hu et al. [18] adopted audio samples, and the RNN model captured temporal patterns from the sequence data samples. The authors have also used the triplet loss function to model discriminative characteristics of the BD severity. The performance analysis is carried out based on the audio/visual emotion challenge dataset. The outcome suggests the effectiveness of the proposed model. The work carried out by Matsubara et al. [19] focuses on the issue of overfitting and extracting a non-relevant feature for the detection of mood swing disorder. The authors have introduced the deep neural generative system in conjunction with Bayes' rules to determine the subsequent probability of the patient state from the given imaging data.

The work of Cigdem et al. [20] has investigated the impacts of covariates on the classification of BD using structural MRI. Voxel-based morphometry is utilized to assess the morphological differences between BD patients and healthy controls. Fitrati et al. [21] used a backpropagation mechanism to detect the BD using a screening questionnaire. The application of the association rule is considered in the study of Castro et al. [22] to determine the relationship between premenstrual dysphoric disorder and BD. The presented technique is evaluated on a dataset of a cohort of 1099 women. A hypothetical study is conducted in Buttenbender et al. [23]. The authors have considered data generated from mobile devices to identify neuropsychiatric disorders' behavior and support decisions. The entire procedure is carried out in three steps. First, neuropsychiatric disorders are distinguished using dispersion dataset, contextual information based epidemiological profile extraction, and finally, the model is validated against time complexity and scalability. Nunes et al. [24] have suggested a hybrid model that integrates modern and consistent tools using the multicriteria methodology to facilitate the early diagnosis of various psychological disorders, including BD. This study offers a good decision support system, but it is limited in scalability. The work towards the classification of BD type-1 and BD type-2 using data modeling technique is carried out by Lee et al. [25]. The authors have built a complementary diagnostic classifier and Gene Ontology (GO) to support the decision-making and functional analysis. Villasanti et al. [26] have presented mathematical modeling based on the differential equation to analyze mood dynamics. The work of Huang et al. [27] focuses on detecting uni-polar and BD based on the pattern obtained from elicited speech. A latent affective structure model and the spectral clustering technique for the data modeling are developed. For early diagnosis and prospective examinations, Demir et al. [28] attempted to determine the distinct features of brain white

matter in BD. The work of Yashaswini et al. [29] introduces a mechanism of modelling of transition dynamic of mixed mood for controlling the behavioral stimulation of a subject. The study contributes to attain fine information related to dynamics of mood transition.

III. Research Problem

The shortcomings of existing schemes in detecting and distinguishing human bipolar disorder have prompted researchers to attempt data-driven and predictive systems modeling. Based on a literature review, it has been explored that there are still substantial issues in existing approaches that need to be improved with a unique implementation strategy despite many research efforts. This section highlights the critical issues and research gaps identified to benefit the core research area.

- Lack of effective exploratory analysis: It has been identified that much of the existing research works suffers in determining meaningful information from complex time series data. The implications of their implementation and execution are uncertain and limited to explore how ML can be leveraged for patient care.
- Lack of effective benchmarking: The existing data-driven approach still encounters acceptance issues in the healthcare industry, as there is no clear consensus on whether such methods are reliable.
- Lack of optimization: Adoption of CNNs is more common in the literature due to their ability to detect critical features automatically without human supervision. However, it is computationally expensive and lacks the ability for the input data to be spatially invariant.
- Prone to outlier issue: The predictive model used to estimate statistical attributes from a temporal data sample is highly sensitive to outliers because the learning model corrects errors in the antecedents.
- Limited to specific design context: The modeling and implementation of existing data-driven schemes are limited to specific design contexts and may not provide similar performance when different contexts are introduced. The predictive ability of the existing learning models is restricted to the analysis of certain features. None of the research work is exhaustive in this manner; the clinical usefulness of the particular features used to derive these models must be considered.

Concurrently, it has been realized that most existing studies are still in the proof-of-concept stage, with small sample sizes and a lack of sufficient validation of prediction algorithms. Furthermore, the quantity and quality of the dataset limit the performance of the algorithm. It has been observed that current machine learning models are subjected to overfit issue, limited to theoretical validation, and suffers in readiness for clinical decision and implementation for effective patient care. Therefore, there is a need to develop an efficient scheme that can predict the true state of bipolar disorder, thereby facilitating the clinical decision-making process for treating patients without consuming too many computational resources and less prone to overfitting problems. The next section discusses the solution which is meant to address the research problem identified.

IV. Research Methodology

The existing approaches to bipolar disorder are often prone to misdiagnosis, leading to difficulties in treating patients. The prime reason for choosing the proposed system is that an effective diagnosis of bipolar disorder has yet no scientific benchmarked model with higher success rate. Available solutions do exist but they are associated with limitation. Out of various problems in existing system, the proposed system chooses to address the research problems highlighted in Section III. The prime aim of the proposed study is to evolve up with a novel computational approach of detecting the state of bipolar disorder based on the feature obtained from motor activity. The significant contribution of this paper is the development of the novel neural network model, which is inspired by the architecture of the decision tree. Unlike existing approaches, the study aims to provide a better decision-making system in diagnosing mood swings disorder with emotional highs and lows with low time complexity and reduced possibility of overfitting. The schematic architecture of the proposed system is depicted in Fig. 1.

The modeling of the entire system adopts a phase-wise implementation procedure. Firstly, exploratory analysis is carried out to understand the characteristics of the dataset and the requirements related to suitable preprocessing. Based on the analysis, the dataset is temporal data as it consists of both time-series and special attributes. It has also been observed that the dataset contains details of 55 individuals, and among them, 23 are diagnosed with bipolar disorder. Further, preprocessing operation is carried out to execute treatment of missing data. In this process, median imputation is employed to handle the missing data. Afterward, descriptive statistics and curve fitting method-based correlation analysis are carried out from the viewpoint of the feature engineering task. In this phase, an analysis is performed regarding a MADRS score (i.e., clinician-rated scale) indicating the depression level. Based on the feature analysis, a new feature descriptor is created named delta-MADRS to analyze the severity of bipolar disorder in the patient. On the other hand, the obtained features are vectorized using a one-hot encoding technique to make the dataset fit to the proposed learning model to predict the actual state of bipolar disorder [30]. The proposed learning design adopts recurrent characteristics and a decision tree structure to better generalize the pattern from training samples that consist of both time-series and discrete observations.

The above Fig. 2 highlights the methodology of the implementation which is carried out in python environment. It is to be noted that the adopted dataset is not an image but it consist of various information in the form of text and numbers. After giving the patient meta-data information as an input in python script, the program access the time-series data which further consist of condition folder and control folder. Condition folder consists of time series data of the people where a doctor

has diagnosed with bipolar disorder while control folder consists of time series data of the people who are in control group. Control group is those people where doctor has confirmed that they don't have bipolar disorder. This is further followed up by extracting MARD scores which is further subjected to correlation analysis for all the columns of data in MARDS score table. Categorical values are evaluated and machine learning approach is applied. This success rate is assessed by the multiple parameters of accuracy analysis explained in result section of this paper.

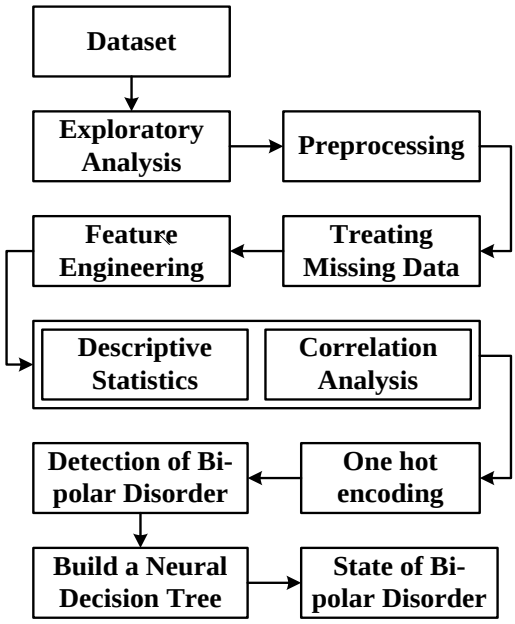


Fig. 1. Schematic Architecture of Proposed System.

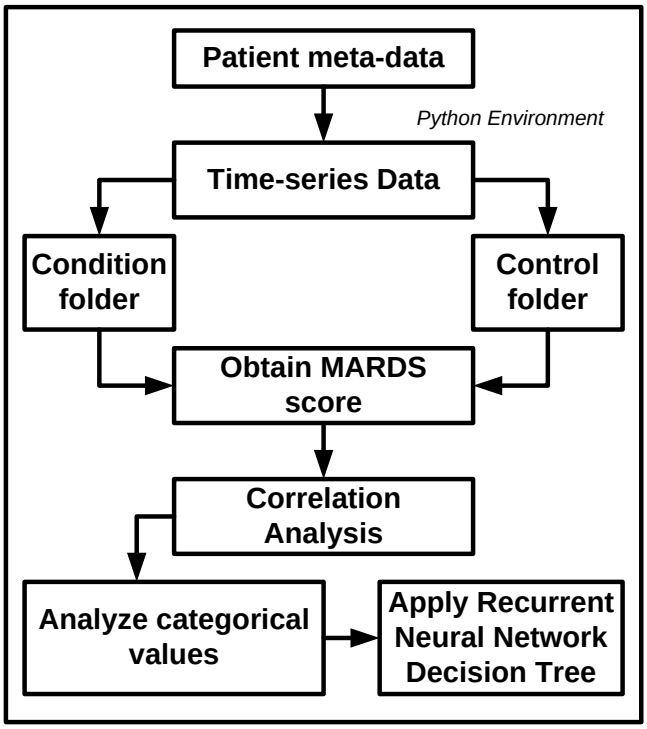


Fig. 2. Methodology for Implementation.

V. System Implementation

The prime target of the proposed system design is to carry out a prediction for the subsequent state of bipolar disorder for the given dataset. This section discusses the algorithm implementation and its respective design execution.

A. Algorithm for Prediction

The core purpose of this algorithm is mainly to carry out a prediction of the subsequent state of bipolar disorder. The core basis of this algorithm is that unpredictable mood is quite challenging to be confirmed as a bipolar disorder. The majority of the existing schemes refer to depression as the indicator of bipolar disorder; however, it is unlikely to confirm this notion. The rationale is that depression could be a state of inactivity, too, and this cannot be concluded as a real state of bipolar disorder. Therefore, the proposed algorithm contributes towards offering a solution to this problem by offering an Actigraph-based monitoring system. This potentially assists in recording the different variants of motor activity of an individual. As the dataset used for the proposed system is available in time series, the prediction could be simplified by applying a unique machine learning approach. The steps of the proposed algorithm are as follows:

Algorithm for Predicting State of Bipolar Disorder

Input: n (elements of the dataset)
Output: *Pred* (predicted outcome)
Start
1. **For** i=1:n
2. **If** n==*err*
3. $n \leftarrow s_{val}$
4. $M=f_1(n)$
5. $C_{val}=f_2(M)$
6. $Pred=f_3(C_{val})$
7. **End**
8. **End**
End

The prime dependency of the algorithm mentioned above is basically a numerical score obtained from MADRS, which is a standard score of rating depression of the subject at the time of observation. This score also assists in offering various statistical numerical values in the form of descriptive analysis. The dataset is preliminarily subjected to preprocessing, followed by the correlation analysis of numerical values. The algorithm implements feature engineering over the data followed by data correlation analysis. The algorithm then finds out the relevance of all the newly obtained data, followed by selecting the subject with reported bipolar disorder. All the time-series data is then loaded, followed by applying a machine learning approach to facilitate the prediction of bipolar disorder. The contribution of this algorithm are viz. i) with a reduced number of iterative operations, the algorithm assists in deploying machine learning operation, ii) the preprocessing carried out offers better elimination of all possible artifacts, iii) a unique feature engineering process is implemented towards obtaining a reliable numerical score of identifying an accurate state of prediction of bipolar disorder. The discussion of algorithmic steps is as follows:

B. Design of Algorithm Execution

This section discusses the step-wise execution of the proposed algorithm that considers a dataset with n elements (Line-1). The proposed system considers the presence of errors *err* within the dataset (Line-2). The algorithm considers the substitution of the errors within the dataset with a substituted value s_{val} (Line-3). The three clauses used for data cleaning purposes are:

- For all the missing data among strings, it is filled with a definitive string.
- For all the missing data among integers, be it discrete or continuous, the empty columns are filled with the median value.
- All categorical data is converted to integers using the ONE HOT encoding technique.

The significance of using a function $f_1(x)$ by considering all the elements *n* of the dataset, which finally leads to the construction of a new matrix *M* (Line-4). The basic idea is to observe the changes appearing within the MADRS score. The advantage of performing this operation is obtaining a standard statistical numerical score that can offer a trend of positive depression ratings. However, there is still a possibility of complexity in the inference obtained from this statistical score in order to understand their connectivity concerning relevance. Therefore, the proposed system carries out correlation analysis using an explicit function $f_2(x)$. Its significance is that it can carry out and assists in graphical-based analysis is further used to find the correlation (Line-5). It is to be noted that this correlation analysis is carried out to assess the relevancy of the MADRS scores with respect to all the categorical values.

The final step of implementing the proposed study is basically applying a machine learning scheme, and a function $f_3(x)$ is developed for this purpose (Line-6). The significance of this function is that the proposed system follows the principle of decision tree architecture. The decision tree's applicability is high in the presence of a maximum number of discrete values compared to the continuous values, which is the pertaining case of the considered dataset. Therefore, the applicability of the decision tree is quite higher in the presence of computed correlated values $C_{val,}$ The significance of this parameter is that it leads to the predicted outcome of *Pred* (Line-6). As the dataset consists of both time series and discrete values, a novel ANN known as Recurrent Neural decision tree is being proposed.

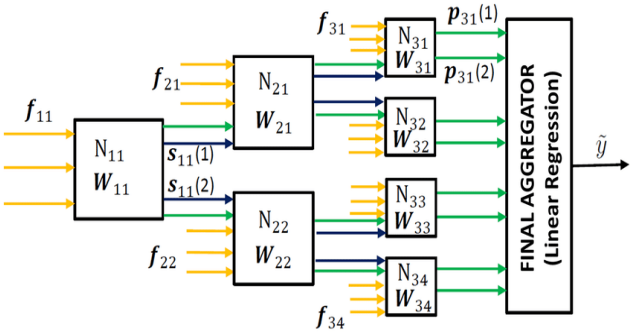


Fig. 3. Architecture of Proposed RNDT.

The architecture exhibited in Fig. 3 exhibits that every node in the decision tree is replaced with 3 neurons which will decide whether the activation should flow towards the right or left. It also shows the presence of input neurons N_{11}, which is further split to N_{21} and N_{22}. Further, the neuron node N_{21} is splitted to N_{31} and N_{32} while neuron node N_{22} is split to N_{33} and N_{34}. Further, a different set of individual weights W_{11}, W_{21}, W_{22}, ….is applied during every processing. This decides the category of the input. However, since this is a regression problem, the output is a continuous value instead of a single value. There is a presence of an aggregator module at the end that uses linear regression to get the continued value of predicted value $P_{31}(1)$, $P_{31}(2)$,….etc. The suitability of linear regression is quite high it can fine-tune the decision trees in case of regression.

VI. Analysis

From the implementation of the proposed scheme discussed in prior section, it is seen that proposed system make use of a simplified identification techniques on the basis of standard scale using psychometric properties. The adopted scheme is found to offer better consistency in its internal operation to ensure the reliability of this scale with higher acceptance. From the performance of trust factor of contents, the proposed scheme ensures that they must be directly linked with core concept of depression. The proposed system make use of different statistical variants of MADRS which is deployed for the purpose of assessing the efficacy of antidepressant trials since long time. The complete analysis is based on the data which is subjected to this scale for evaluating the depressive symptomology which is quite significant indicator for the patient undergoing antidepressant therapies. It doesn't consider any form of somatic symptoms but it essentially emphasizes over psychological symptoms associated with depression which is characterized by pessimistic thoughts, tension, and sadness. The next section discusses about the result being accomplished while the clinical inference of this usage of proposed system can be analyzed by adopting standards of this scale. The proposed system uses a scale that compromises of ten different items with lower and higher value of scale being 0 and 6 respectively. With accomplishment of higher score, the patient can be declared to be in depressive mode in higher degree. If the cumulative value of the score reside in between 0 and 6 than the patient is considered lack of depression symptoms, if the score is between 7 and 19 than it represents milder degree of depression, if the score is between 20 and 34 than it indicates moderate depression, while the greater score of 35 is considered to have severe depression. The critical case of depression is said to occur when this score is more than or equal to 60.

Therefore, for an effective analysis with respect to granularity, it is essential to obtain the descriptive analysis of this score where score effectiveness can be assessed with respect to different variant of error observation (mean, root mean, and absolute). This inference is used as a back end logic to assess the proposed computational model in order to investigate the degree of bipolar disorder in a patient. The section discusses this with obtain numerical and graphical outcomes.

VII. Result Analysis

This section discusses the results obtained from implementing the algorithm elaborated in the prior section. As stated in the prior section, that algorithm tends to predict the state of action as the prime indicator of the depression level of the subject. The illustration of this section is discussed with respect to the environment that is considered for an assessment, highlighting the results being achieved and a discussion of the results.

A. Assessment Environment

The development of the proposed system is carried out in a Python environment considering a normal windows machine. The implementation is carried out on the depression dataset [31]. The dataset consists of a metadata file as well as two forms of data, i.e., conditional data and controlled data. The former one is a time-series data of the subject who is positively diagnosed with a bipolar disorder by the doctor, while the latter one is also a time-series data that are negatively confirmed to have a bipolar disorder. The next part of the assessment environment is about using MADRS. As bipolar disorder represents a high amount of mood swings, the difference in the MADRS at the beginning of the observation is much different from that during the beginning. The measure of difference also indicates the severity of bipolar disorder as well. Hence a new column called ΔMADRS is created, which is the difference between the MADRS1 and MADRS2.

B. Results Obtained

In order to carry out an effective analysis, the proposed system is compared with the conventional decision tree with respect to R2 score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) as exhibited in Fig. 4 to Fig. 7, respectively. The outcome eventually showcased that the proposed Neural Decision Tree system offers better performance than the conventional decision tree.

TABLE I. Descriptive Analysis Tabulation

Item	DAYS	MADRS1	MADRS2	Δ MADRAS
count	23	23	23	23
mean	12.65217	22.73913	20	-2.73910435
Std	2.773391	4.797892	4.729021	3.968253466
Min	5	13	11	-13
25%	12.5	18.5	16	-5
50%	13	24	21	-3
75%	14	26	24.5	0
Max	18	29	28	4

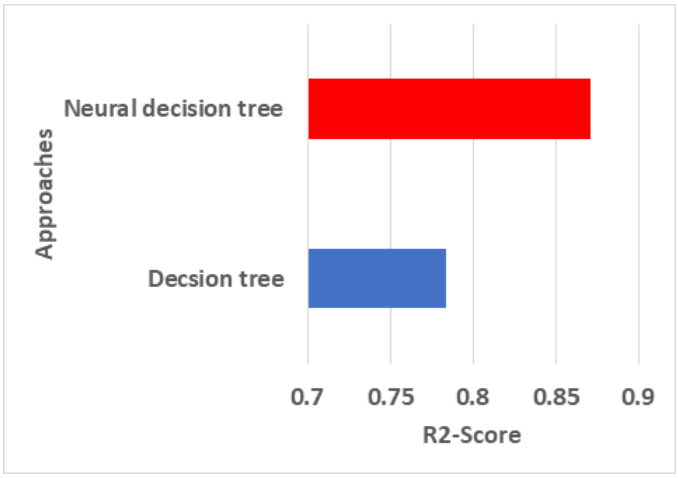


Fig. 4. Comparative Analysis of R2 Score.

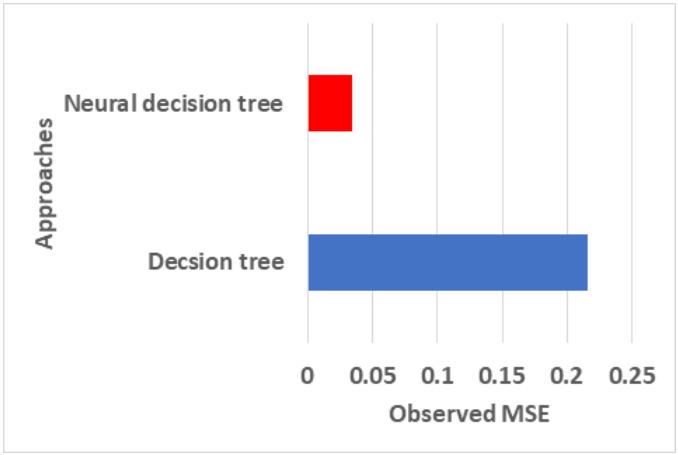


Fig. 5. Comparative Analysis of MSE.

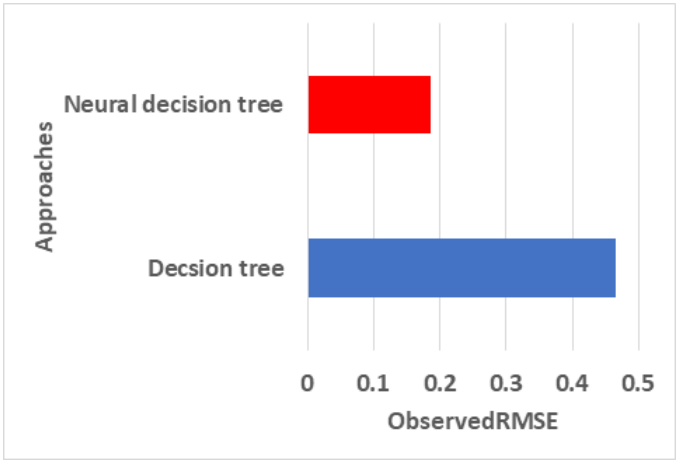


Fig. 6. Comparative Analysis of RMSE.

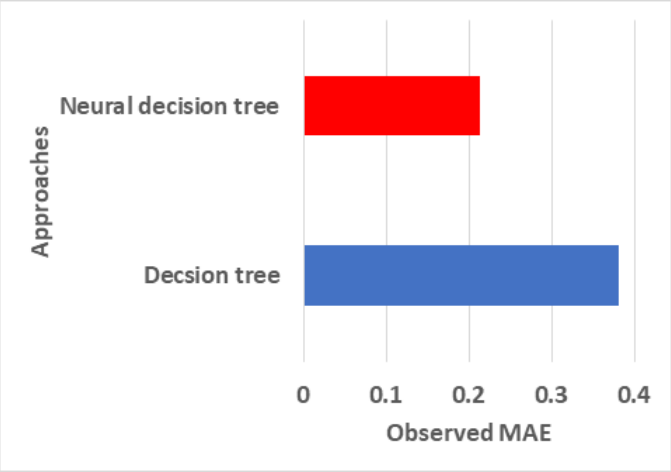


Fig. 7. Comparative Analysis of MAE.

C. Result Discussion

The analysis of the proposed system's outcome initiates from analyzing the MARD score. The numerical outcome exhibited in Table I clearly indicates that more than 75% of the diagnosed people with the MARDS score deteriorated during the observation. The outcome eventually indicates that the MADRS score reduces up after therapy compared to the prior MADRS score before therapy. The next set of outcomes is obtained from the correlation analysis for all the numerical values in Table I. The outcome indicates that the MARDS1 and 2 are correlated to each other. The above graphs use the curve fitting method to indicate the correlation. Based on these, more conclusive remarks can be obtained over the dataset as follows:

- Age: People with age-group of 45-49 as well as 35-39 are more prone to depression.
- Gender: Women are found to be more susceptible for bipolar disorder compared to man.
- Type of Bipolar Disorder: Type-2 bipolar disorder is found to be more predominant in contrast to any other type.
- Mood Swings: The analysis found that the patient often does not indicate sadness, making the diagnosis harder. The bipolar disorder only means frequent mood swings but not prolonged sadness. Due to these mood swings, it becomes difficult to predict the behaviors of the patient. This proves the scope of the study that since it is difficult to predict, a machine-learning algorithm has been used to help the therapist.
- Type of Admission: Analysis shows that bipolar patients are seldom admitted to the hospital most of the time; they are outpatients who can be treated like normal patients for any other disorder.
- Education: The lesser the person is educated, the more likely the person gets bipolar disorder.
- Marriage: Bipolar disorder has nothing to do with marital status.

VIII. Conclusion

There have been various myths about a proper diagnosis of a patient with bipolar disorder. In various analyses, it is a myth to associate sadness or a specific state of mind to bipolar disorder. The limiting factor of this paper could be that it doesn't consider any form of somatic symptoms as it uses MADRS scheme. It doesn't affect the outcome as the proposed system contributes towards connecting activity-based state with depressive scale as a novel approach. However, the proposed system proves that adopting an activity-based state is the precise means to perform an analysis. The novelty introduced by the proposed study is as follows:

- Unlike any existing studies on bipolar disorder, the proposed scheme doesn't directly consider the dataset as it is rather, it performs a set of operations to make it more suitable for analysis,
- The proposed system makes use of Actigraph for considering the motor activity of a subject, thereby considering time-series based data suitable for predictive analysis,
- The complete analysis is carried out considering all the meta-data present in the dataset with respect to days of observation, age, type of diagnosis, melanch, type of admission, marriage, employment, etc.,
- The proposed system introduces three sequential processes for data cleaning in order to find all the missing data as well as the transformation of the categorical data,
- The model harnesses the standard scale of MADRS score for accounting for the fluctuation in the mood swing by exploring the difference in the individual progressive scores, vi) contextual accuracy in the data inference is obtained by applying contextual analysis,
- Usage of a recurrent decision tree that hybridizes a recurrent neural network with a decision tree for making precise predictive outcomes for bipolar disorder.

The quantified outcome of the proposed study are as follows: The proposed scheme offers approximately 50% of betterment in R2 score compared to conventional decision trees, approximately 80% reduced value of MSE, 62% reduced value of RMSE, and 41% reduction of MAE in comparison to conventional decision tree approach. Hence adoption of proposed neural decision tree significantly assists in modelling various collective activity-based data with each pattern arranged in form of decision tree with less dependency. This significantly assists in achieving faster as well as simpler computational efficiency while executing the model.

The future work of the proposed study could be in direction of further optimizing the model with more constraint inclusion. This could be further achieved by non-linear constraint modelling where along with activity, various other non-linear non-activity based parameter could be opted. The future work would also be in direction towards inclusion of multi-objective function to cater up cost effective modelling demands.

REFERENCES

[1] Paris, Joel. "Mood disorders and personality disorders: Simplicity and complexity." In Borderline Personality and Mood Disorders, pp. 3-9. Springer, New York, NY, 2015.

[2] Singh, Meharban. "Compulsive digital gaming: an emerging mental health disorder in children." The Indian Journal of Pediatrics 86, no. 2 (2019): 171-173.

[3] Blackwood, Douglas HR, Ben J. Pickard, Pippa A. Thomson, Kathryn L. Evans, David J. Porteous, and Walter J. Muir. "Are some genetic risk factors common to schizophrenia, bipolar disorder and depression? Evidence from DISC1, GRIK4 and NRG1." Neurotoxicity Research, vol.11, no. 1, pp.73-83, 2007.

[4] Gilkes, Melissa, T. Perich, and T. Meade. "Predictors of self-stigma in bipolar disorder: Depression, mania, and perceived cognitive function." Stigma and Health, vol.4, no. 3, 2019.

[5] Tondo, Leonardo, Gustavo H Vazquez, and Ross J Baldessarini. "Depression and mania in bipolar disorder." Current neuropharmacology, vol.15, no. 3, pp.353-358, 2017.

[6] Koukopoulos, A., Reginaldi, D., Tondo, L., Visioli, C. and Baldessarini, R.J., "Course sequences in bipolar disorder: depressions preceding or following manias or hypomanias." Journal of affective disorders, vol.151, Iss.(1), pp.105-110, 2013.

[7] Yashaswini, K. A., and Shreyas Rao. "Bipolar Disorder: A Pathway Towards Research Progress in Identification and Classification." In Computer Science On-line Conference, pp. 205-214. Springer, Cham, 2020.

[8] Yashaswini K.A., Saxena A.K., "Novel Classification Modelling for Bipolar Disorder Using Non-verbal Attributes for Classification". In: Venugopal K.R., Shenoy P.D., Buyya R., Patnaik L.M., Iyengar S.S. (eds) Data Science and Computational Intelligence. ICInPro 2021. Communications in Computer and Information Science, vol 1483. Springer, Cham., 2021 https://doi.org/10.1007/978-3-030-91244-4_19.

[9] Vasu, V. and Indiramma, M., 2020, September. "A Survey on Bipolar Disorder Classification Methodologies using Machine Learning". IEEE International Conference on Smart Electronics and Communication , pp. 335-340, 2020.

[10] Phillips, M.L. and Kupfer, D.J., "Bipolar disorder diagnosis: challenges and future directions", The Lancet, vol.381, Iss.9878, pp.1663-1671, 2013.

[11] Ceccarelli, F., Mahmoud, M. Multimodal temporal machine learning for bipolar disorder and Depression Recognition. Pattern Anal Applic, ISSN 1433-754, 2021. DOI: https://doi.org/10.1007/s10044-021-01001-y.

[12] Sánchez-Morla, E.M., Fuentes, J.L., Miguel-Jiménez, J.M., Boquete, L., Ortiz, M., Orduna, E., Satue, M. and Garcia-Martin, E., "Automatic Diagnosis of Bipolar Disorder Using Optical Coherence Tomography Data and Artificial Intelligence", Journal of Personalized Medicine, vol.11, Iss.(8), p.803, 2021.

[13] Mateo-Sotos, J., Torres, A.M., Santos, J.L. et al., "A Machine Learning-Based Method to Identify Bipolar Disorder Patients". Circuits Syst Signal Process, 2021. https://doi.org/10.1007/s00034-021-01889-1.

[14] Chen, ZhiHong, T. Yan, E. Wang, H. Jiang, Y. Tang, X. Yu, J. Zhang, and C. Liu. "Detecting abnormal brain regions in schizophrenia using structural MRI via machine learning." Computational intelligence and neuroscience, 2020.

[15] S. Qiu, Q. Yue, F. Zhu, and K. Shu. "The Identification research of bipolar disorder based on CNN." In Journal of Physics: Conference Series, IOP Publishing, vol.1168, no. 3, p. 032125, 2019.

[16] Li, Z., Li, W., Wei, Y., Gui, G., Zhang, R., Liu, H., Chen, Y. and Jiang, Y., "Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls". Computerized Medical Imaging and Graphics, 89, p.101882, 2021.

[17] Huang, Kun-Yi, Chung-Hsien Wu, and Ming-Hsiang Su. "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses." Pattern Recognition, vol.88, pp.668-678, 2019.

[18] Du, Z., Li, W., Huang, D. and Wang, Y., 2018, October. Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (pp. 23-30), 2018.

[19] T. Matsubara, T. Tashiro and K. Uehara, "Deep Neural Generative Model of Functional MRI Images for Psychiatric Disorder Diagnosis," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 10, pp. 2768-2779, Oct. 2019, doi: 10.1109/TBME.2019.2895663.

[20] O. Cigdem et al., "Effects of Covariates on Classification of Bipolar Disorder Using Structural MRI," Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019, pp. 1-4, doi: 10.1109/EBBT.2019.8741586.

[21] D. Fitriati, F. Maspiyanti and F. A. Devianty, "Early Detection Application of Bipolar Disorders Using Backpropagation Algorithm," 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), pp. 40-44, 2019, doi: 10.23919/EECSI48112.2019.8977102.

[22] G. Castro et al., "Applying Association Rules to Study Bipolar Disorder and Premenstrual Dysphoric Disorder Comorbidity," IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), pp. 1-4, 2018 doi: 10.1109/CCECE.2018.8447747.

[23] P. C. Büttenbender, E. G. d. A. Neto, W. F. Heckler and J. L. V. Barbosa, "A computational model for identifying behavioral patterns in people with neuropsychiatric disorders," in IEEE Latin America Transactions, vol. 20, no. 4, pp. 582-589, April 2022, doi: 10.1109/TLA.2022.9675463.

[24] L. C. Nunes, P. R. Pinheiro, M. C. Dantas Pinheiro, M. Simão Filho, R. E. Comin Nunes and P. G. C. Dantas Pinheiro, "Automatic Detection and Diagnosis of Neurologic Diseases," in IEEE Access, vol. 7, pp. 29924-29941, 2019, doi: 10.1109/ACCESS.2019.2899216.

[25] C. -Y. Lee, J. -H. Zeng, S. -Y. Lee, R. -B. Lu and P. -H. Kuo, "SNP Data Science for Classification of Bipolar Disorder I and Bipolar Disorder II," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2862-2869, 1 Nov.-Dec. 2021, doi: 10.1109/TCBB.2020.2988024.

[26] H. G. Villasanti and K. M. Passino, "Modeling and Analysis of Mood Dynamics in the Bipolar Spectrum," in IEEE Transactions on Computational Social Systems, vol. 7, no. 6, pp. 1335-1344, Dec. 2020, doi: 10.1109/TCSS.2020.3028205.

[27] K. -Y. Huang, C. -H. Wu, M. -H. Su and Y. -T. Kuo, "Detecting Unipolar and Bipolar Depressive Disorders from Elicited Speech Responses Using Latent Affective Structure Model," in IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 393-404, 1 July-Sept. 2020, doi: 10.1109/TAFFC.2018.2803178.

[28] A. Demir, M. Özkan and A. M. Uluğ, "A Macro-Structural Dispersion Characteristic of Brain White Matter and Its Application to Bipolar Disorder," in IEEE Transactions on Biomedical Engineering, vol. 68, no. 2, pp. 428-435, Feb. 2021, doi: 10.1109/TBME.2020.3002688.

[29] Kumar, Yashaswini Kunjali Ajeeth, and Aditya Kishore Saxena. "Stochastic modelling of transition dynamic of mixed mood episodes in bipolar disorder." International Journal of Electrical & Computer Engineering vol.12, pp.2088-8708, no. 1, 2022.

[30] Y. H. Liu, Python Machine Learning By Example, Packt Publishing, ISBN: 9781783553129, 178355312X, 2017.

[31] A. Jan, H. Meng, Y. F. B. A. Gaus and F. Zhang, "Artificial Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions," in IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 668-680, Sept. 2018, doi: 10.1109/TCDS.2017.2721552.

Objective Type Question Generation using Natural Language Processing

G. Deena[1], Dr. K. Raja[2]
Research Scholar, Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology, Rajiv Gandhi Salai, Chennai, 600119, India[1]
Department of Computer Science and Engineering, SRM Institute of Science and Technology
Bharathi Salai, Chennai, 600089, India[1, 2]

***Abstract*—Automatic Question Generation (AQG) is a research trend that enables teachers to create assessments with greater efficiency in right set of questions from the study material. Today's educational institutions require a powerful tool to correctly assess learner's mastery of concepts learned through study materials. Objective type questions are an excellent method of assessing a learner's topic understanding in learning process, based on Information and Communication Technology (ICT) and Intelligent Tutoring Systems (ITS).Creating a set of questions for assessment can take a significant amount of time for teachers, and obtaining questions from external sources such as assessment books or question banks may not be relevant to the content covered by students during their studies. This proposed system involves to generate the familiar objective type questions like True or False, 'Wh', Fill up with double blank space, match the following type question have generated using Natural Language Processing(NLP) techniquesfrom the given study material. Different rules are created to generate T/F and 'Wh' type questions. Dependence parser method has involved in 'Wh' questions. Proposed system is tested with Grade V Computer Science text book as an input. Experimental result shows that the proposed system is quite promising to generate the amount of objective type assessment questions.**

Keywords*—Intelligent tutoring system; true or false; dependency parser; natural language processing; question generation

I. INTRODUCTION

In academics, an assessment is critical to determine the skill level of a learner. Question generation has emerged as a new area of study in the fields of traditional and in online education and natural language processing [1, 2, 3][4,5]. Research has been done to improve the effectiveness of automatic question generation in recent years. Acquiring the knowledge is the ultimate goal of an educational system, but assessment or evaluation is the final goal in learning [6,7,8]. Evaluation process needs questions to measures the learner's topic knowledge [9, 10,11].

Automatic Question Generation (AQG) is the process of automatically creating syntactically valid, semantically correct, and relevant questions from a variety of inputs in the form of text, a structured database, or even a knowledge base as given in [12,13]. AQG is a perfect fit for a variety of areas, including Massive Open Online Courses (MOOC), Chatbot systems for customer engagement and interaction and healthcare system are a few examples of automated help systems as given in [14]. Computerized practices can be produced better than standard specialists, while still maintaining quality.

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that studies how machines and humans communicate using natural language. The goal of NLP is to identify, comprehend, grasp, and interpret the recorded or text data in human languages, whether in English or in another language, in a manner that is beneficial. Subjective and Objective type questions are two primary methods in traditional and digital educational system [15]. Due to the benefit of rapid and real-time examination, objective type test items are gaining great significance in an Intelligent Tutoring System (ITS) and in dynamic classroom learning. Since there are many uses in conversational AI systems and education, interest in the topic has grown.

Objective types questions are termed as Cloze Type (CT) Question which can be answered in short or in one word. They are very useful for learning specific ideas, and they are frequently employed in school level. However, these questions may be manually prepared, and this necessitates significant understanding on the topics being prepared. The individual or the teacher must examine each sentence of the section and determine whether or not each statement is capable of generating a question. A valid question can be derived from a given sentence only if it includes information that contributes to the learner's knowledge gain. Input is a text file and an output is a generated objective type True or False and 'Wh' type questions using NLP techniques. It is necessary to select the proper response from a list of possibilities, or finish a statement with the use of an appropriate sentence fragment or word. Commonly utilized objective exam items include True or False, Multiple-Choice Questions, and Fill-in-the-blank questions [16].

In "Wh" is investigating where someone will be responsible for generating item pools with items like "What," "When," "Where," "Who," and "Which." Researchers have been focused on automatic CQ creation for the past decade [17,18]. AQGcan support the teacher and the learner entity. Additional practise can be given as self-assessment, is provided by AQG to learner. Alternatively, AQG can also aid teachers in the laborious process of deciding on the type of questions they will pose to the learner. Also, AQG can assist online learners in creating questions from random text. The

concept of effective questioning is a strategy to help students pay attention to their learning goals. As well, to ensure its effectiveness in integrating Bloomslower level and high-level taxonomy the questions are required as mentioned in [19].

The proposed system focuses on developing different types of objective questions like Fill-Up with doubleblank questions, True or False questions, Match the following and short answer questions to assess a learner's learning gap. The majority of phrases in a text do not lend itself to the generation of high-quality queries. As a result, the informative sentence selection assignment has done to generate questions from those sentences. As the continuation of our previous work, the informative sentences are selected using Latent semantic Analysis given in [20]. Answer key is the top most word which yield the relation between the sentences and those answer keys are replaced by the blank space.

In True or False type questions, the answer key has identified from the Part-of-speech tag. Tag can be the auxiliary verbs, cardinal number, adjective, negative words of the sentence Based on these tag value and predicting the next word involved in falsified sentence is generated as true of false questions using Natural Language Processing. The sentences will be falsified by changing the tag word.

Fill-up with double blank question, the informative sentence and important word of the sentence is identified using Latent semantic analysis. The important word will be the answer key of the sentence.

In terms of short answer questions, the dependency parser and Named Entity Recognizer (NER) are involved to frame the generate questions. NER will give the Person Name, Place, Organization, and Location from the sentence. The dependency parser gives the "SUBJ, NMOD, DOBJ, DET, CASE, NSUBJ, CC" of the sentence. Both the NER and dependency parser combined together to generate the short answer questions.

In match the following type questions, the abbreviated words are selected from the corpus and consider as input in the first column and the randomly the expansions are given at the right side in the second column as an output.

Evaluation of the proposed system is done through grade V Computer science book of Central Board of Secondary Education. The paragraph from the book is given as input to the system. The pre-processing has been applied to clean the input. Later, the system generates the different type of question and the quality of the system is manually verified by the school teacher. The cardinality and validity are the two parameters to measure the performance of the proposed system. The generated questions were tested with the students of grade Vto ensure the understandability of the questions. In general, the experiment observed that 92% of validity has been achieved.

II. Related Work

In recent years, development in an automatic question generation system seems to have become a hot research topic. There are a number of different systems and methodologies and some of which use distinct methods and practices to generate questions automatically. The tasks focus on specific methods that use sentences or phrases out of unstructured material to generate questions for students, which are used to assist in their knowledge assessment.

In numerable question generation systems have been suggested previously like Fill up the blanks, 'Wh' questions to assess the skill of a learner. The whole work has been divided into different phases like sentence identification, extracting the important keyword and question formation. In the existing methods, different methods were used to extract keywords the appropriate keyword in the sentence is replaced by the blank space to generate the Fill-up-based questions. In the proposed system, the informative sentence and keywords are extracted using Latent Semantic Analysis. and more than one keyword in the sentence is replaced by the blank space to generate the fill up type questions.

Sheetal Rakangoret.al, [21] suggested the true or false question by adding the "not", "no" and "never" in the affirmative sentences. The affirmative sentence is generally positive statement which gives sufficient information to the reader. Those sentences will not generally carry any negative words in the sentence. According to the author, the part-of-speech for a sentence is identified and add 'no', 'never', 'not' before the auxiliary verbs. This system converts the affirmative sentence into negative sentences. In the proposed system, numbers of rules are created to generate the questions using NLP techniques.

In the short objective type questions, many researchers have focused to develop factual question based on AUTOQUEST as suggested in [22] that was the first method to generate automatic factual question. Syntactic principles were first used to generate questions, and then some type of statistical assessment of goodness was done in [23].

The other works discuss in this chapter are concerned with finding questions in unstructured text sentences or sequences of sentences. Here, lets describe a few approaches to generate the objective type questions generation that are available in the literature. In Section 3, the proposed system true or false type questions were described, Fill up the double blank type question has described in Section 4, Match the following carried in Section 5, short question answer questions mentioned in Section 6, the experimental result shown in Section 7 and conclusion mentioned in Section 8.

III. Proposed System

A. True or False

Different type of objective type questions is in practice to assess the learner. True or False, Match the following, fill up the blanks with double blank, Wh-type or One word Question Answer are the variety of question types. True or False (T/F) is the quickest process of measuring the learner. T/F can be resolved by returning a binary value of "True" or "False." Once the learner grasps the subject clearly, the T/F questions are answered as part of the lowest level of the bloom taxonomy. Proposed system works on the basis of sentence tokenization, word tokenization and Part-of-Speech (POS) identification. The affirmative sentence has been involved in question generation.

B. Sentence Tokenization

The given corpus consists of sequence of sentences which can be tokenized into individual sentence. Each sentence will be analysed separately to generate the objective questions. In the current system, to generate the True or False(T/F), the sentence tokenizer is used to extract individual sentences using Natural Language Processing(NLP) techniques. Input paragraph is tokenized into individual sentence. Individual sentences are analysed to find any affirmative sentences, holding any cardinal value, any negative sentences.

C. Keyword Identification

Once the passage contains statements about factual information then question can be raised in the form of True of False. In the proposed system, the given corpus is tokenized into sentences using the sentence tokenizer followed by the word tokenizer to tokenize into individual words. The part-of-speech (POS) tag set is applied that has the ability to label the words as noun, adjective, adverb, and verb. Keywords are identified by the POS are also known as tokens.

D. Rules for True or False Questions

In the proposed system, the affirmative sentences are changed into negative sentence, identify the antonyms to adjective words, changing the cardinal value using random function, using GPT-2, replace the last word with options and in terms of the negative sentence remove the negative words to generate True or False questions for the given corpus. A rule has been created for mentioned types of ways to generate the T/F Questions.

A positive or an affirmative form is employed to convey the validity or reality of a fundamental proposition of the sentence. There will be no negative phrases like “not”, “never”, “none” in a positive sentence which represents the truth of the sentence. This sentence is factually correct and gives some information to the learner.

A negative sentence is a statement that something is false or incorrect. To cancel the sentence's credibility, a negative adverb must be added. Affirmative phrases show their falsity through the negative statement. Don't negate a negative sentence rather make it to positive sentence.

Rule 1: Negate the affirmative statement

Affirmative sentences are the sentences that have positive meaning. An auxiliary verb phrases assists another verb in its tense or in voice formation. Auxiliary verbs are also referred as assisting verbs. These are called as helping verbs to the words in the sentence. Auxiliary verbs are like “am, are, is, was, were, can, could, may, have”. Perform the steps to generate the falsified sentence.

1) Check if the sentence in the corpus has any auxiliary verb, then add “Not” before the auxiliary verb to negate the sentence.

2) The affirmative sentence has become falsified sentence as True or False questions.

3) Print the question.

Let us consider the samples input sentence from the Grade V Computer Science Book as in Table I.

Rule 2: Adjective word replacement

Another possibility is to change the meaning of a sentence is by setting the antonyms to the label adjectives. The antonyms of the adjectives set from WordNet. WordNet is a lexical database built for natural language processing in English, specifically. English Lexical Database (ELD) is a dictionary which consists of different kind of similar meaning and opposite meanings to a single word. Synsets instances are groups of words that can be used to define the same concept. The Nouns, Verbs, Adjectives, and Adverbs are classified according to their cognitive and linguistic relationships of a word. The given steps follow to find the antonyms of the word.

1) Apply POS to extract the adjective tag of a word from the sentence.

2) Apply the lemmatization to determine its word's root.

3) Apply the antonym function to find the opposite meaning of the root word to alter the meaning of a statement.

4) Generate the falsified questions.

The adjective of the word is replaced by the antonyms of that word looked up from the Wordnet as given in the Table II.

Rule 3: To replace the cardinal number

This rule is simple to alter the original meaning of a statement by altering its cardinal value. Find whether any the number is present in the sentence using the POS. Falsify the given statement by using the random function to change the cardinal value. The sample was given in the Table III.

1) Check if the tag of word is ‘cardinal’ then apply the random function to change the value of the cardinal.

2) The new cardinal value makes the sentence into falsified.

TABLE I. NEGATED AFFIRMATIVE SENTENCES

Auxiliary Verb	Input sentence	Falsified Sentence
was	ENIAC was the first fully electronic digital computer	ENIAC was not the first fully electronic digital computer
was	The ENIAC was developed by John Presper Eckert	The ENIAC was developed by John Presper Eckert
were	First Generation computers were hugely expensive to build	First Generation computers were not hugely expensive to build
was	EDSAC was the first electronic computer that used stored programs.	EDSAC was not the first electronic computer that used stored programs.
was	EDSAC was the first electronic computer that used stored programs.	EDSAC was the first electronic computer that used not stored programs.
have	Second Generation computer have used transistors	Second Generation computer have not used transistors

TABLE II. ADJECTIVE WORD REPLACEMENT

Adjectives	Input sentence	Falsified Sentence
large	First Generation very large in size	First Generation very small in size
expensive	First Generation very expensive	First Generation very cheap
slow	First Generation operating speed was quite slow	First Generation operating speed was quite fast
slow	First Generation operating speed was quite slow	First Generation operating speed was quiteaccelerate.
cheaper	Third generation were cheaper than second generation	Third generation were expensive than second generation

TABLE III. ALTERNATIVE OF CARDINAL NUMBER

Cardinal	Input sentence	Falsified Sentence
360	IBM 360 examples of third generation	IBM 362 examples of third generation
1948	EDVAC was completed in the year 1948	EDVAC was completed in the year 1957
6000	EDSAC weighted approximately 6000 kilograms.	EDSAC weighted approximately 6006 kilograms.
1946	ENIAC was developed by John Mauchly in the year 1946.	ENIAC was developed by John Mauchly in the year 1955.
1401	IBM 1401 are example of second generation	IBM 1403 are example of second generation

Rule 4: To predict the last word

To falsify the statement, another possibility is predicting the last word of the sentence. The last noun labelled word is extracted separately from the sentence. The extracted word is now then termed as keyword. Predicting the subsequent word is one the task in NLP activities.GPT-2 is an open-source Artificial Intelligence with 40GB of text is analysed to train the prediction algorithm. Generative Pre-trained Transformer-2 (GPT-2) is the leveraged transformer in both unsupervised and supervised text representation. To falsify the sentence, perform the steps to predict the alternatives of the keyword. The implemented samples rules were given in Table IV.

1) Split the actual sentence into two halves by using rsplit().

2) Rsplit() function to split the last word from the sentence.

3) Import GPT-2 from next_word_prediction.

4) Invoke predict_next() with an rest of the sentence as argument.

a) It produces the possibilities of the left-out word

b) Select any of those word to generate the falsify statement and print.

Rule 5: Converting Negative Sentence into Positive Sentence

A negative sentence is one that declares something to be false and it indicates that a specific statement or circumstance is incorrect. List of sample negative words are given in Table V. To generate the true or false question lets convert the negative sentence to positive sentence as affirmative by following the steps,

1) Check if any of the negativity is present in the sentence, and then eradicate them to convert the sentence as affirmative.

2) Affirmative sentence is generated.

The sample of the removing the negative words were given in the Table VI.

TABLE IV. PREDICTING THE LAST WORD BY GPT-2

Last word	Input sentence	Falsified Sentence
time	Microcomputer is a small computer that is used by one person at a time	Microcomputer is a small computer that is used by one person at a desk.
PC	Desktop computers are also called as PC	Desktop computers are also called as windows
2000	ENIAC was developed by John Mauchly in the year 2000	ENIAC was developed by John Mauchly in the year 2000
Slow	First Generation operating speed was quite slow	First Generation operating speed was quite good
expensive	First Generation very expensive	First Generation very early

TABLE V. NEGATIVE WORDS

no	not	none	Never	nothing
don’t	can’t	won’t	wasn’t	isn’t

TABLE VI. REMOVAL OF NEGATIVE WORDS

Negative words	Input sentence	Falsified Sentence
not	Mano is not a professional singer.	Mano is a professional singer.
never	The coffee shop is never open	The coffee shop is never open
won't	They won't be eating	They won't be eating
don't	Mithul don't play keyboard	Mithul play keyboard
aren't	They aren't coming to school in pandemic situations	They are coming to school in pandemic situations

E. Fill up the Double Blank based Question

In the work given by author Smith [24], Karamaniset al [25] and Mitkov et al [26] pick appropriate keywords by determining frequency of terms and running regular expressions on nouns. Find the sentences with that keyword in them and their methods produce one-word gap-fill questions. In the continuation of our previous work [27], the important sentence and the keyword was identified through Latent Sematic Analysis (LSA), this method produce the hidden relation between the sentences and the words. To select the keywords, the proposed system fixes the threshold value between 1.5 and 2.0 and those above the value were identified as most important keyword involved in generating the Fill up based assessment questions.

Sentences with two keywords involved in generating the double blank based questions, where the keywords are replaced by the blank space. In the existing system, single keyword is replaced by the blank space. Keyword may be Noun related words extracted using TF-IDF or by keyword extractor. In the proposed system, the LSA used to select the keyword it can be Noun, Verb which has hidden relation between the words and sentences. This double blank fill-based question gives promising result to determine the skill level of a learner. The samples were given in Table VII.

F. Match the Following Questions

In the proposed system, to generate the Match the Following type questions, abbreviation and their expansion, definition type words are taken as input from the given input file and maintained in the .csv file, where the abbreviation words are selected as the keyword. Each abbreviation has the expansion. The table is constructed with the row and column. The keywords are the abbreviated words and define type words, the keywords are the first column and the respective expansion of the keywords and the explanation of the define words are shuffled using shuffle method and prearranged at the second column. Now the keywords have wrong abbreviation in the nearby column. This set is finally made as match the following questions. As a result, the match the following type question are generated and maintained in the table. As part of the objective type question, the learner will take up the assessment to test the skill level. The sample of these questions are executed in jupyter notebook and given in Fig. 1.

G. One Word Question and Answer

"Wh" questions are the type of questions to answer about specific qualities, times, places and people. Like 'Wh'-type question, we generate one word question and answer to determine the learning skill level of a learner. Dependency Parsing is the process of analysing a sentence's grammatical structure and identifying related words, as well as the nature of the link between them using computer programmes. It enables us to construct a parsing tree based on the tags used to determine the relationship between the words in the sentence, rather than on any Grammar rule as with syntactic parsing, which provides a great deal of flexibility even when the order of the words changes.

In the existing system, the Named Entity Recognizer (NER) has used to generate the Who, What, When, Where type questions. Always it's not recommended to generate "Wh" questions. In the existing methods, only one word has been extracted using POS, NER, and dependency parser to generate questions. In the proposed system, the spacy Matcher, Dependency Parser and NER combined together to generate the assessment questions. In some case, two words.

1) Matcher: In spacy, rule-based matching enables you to define your own criteria for locating and extracting words and phrases from a text. To match the individual word or text the matcher is the best option and is easy to create. Here, let set the Matcher's vocabulary to the language of the spacy model in Natural Language Processing (NLP) and maintain a list of dictionaries for the pattern of the type which really required for the sentence. Each dictionary contains the match pattern for a token.

Menu	Definition
ENAIC	Universal Automatic Computer
EDSAC	Electronic Delay Storage Automatic Computer
EDVAC	Artificial Intelligence
UNIAC	Electronic Discrete Variable Automatic Computer
AI	Electronic Numeriacal Integrator and computer

Fig. 1. Match the Following Type Questions.

TABLE VII. FILL UP WITH DOUBLE BLANK –QUESTIONS

S. No	Fill up the double blank questions	Keyword
1	Expansion of _____ was termed as Electronic _____ Integrator and Computer	ENIAC Numerical
2	Expansion of ENIAC was termed as _____ Numerical _____ and Computer	Electronic Integrator
3	Expansion of ENIAC was termed as Electronic _____ Integrator and _____	Numerical Computer
4	Expansion of _____ was termed as Electronics Numerical _____ and Computer	ENIAC Integrator
5	It was the first _____ electronic _____ computer	Fully Digital
6	It was the first fully _____ digital _____	Electronic Computer
7	It was the first fully electronic _____	Digital Computer
8	_____ was an example of first-generation _____	ENIAC Computer
9	The ___ could add two large numbers just in 200 _____, much higher speed than its predecessors	ENIAC Microseconds
10	The ENIAC could add two _____ number just in 200 microseconds, much higher speed than its predecessors	Large Microseconds
11	The _____ could add two large _____ just in 200 microseconds, much higher speed than its predecessors	ENIAC Numbers
12	The ENIAC could add two large _____ just in 200 _____ much higher speed than its predecessors	Numbbers Microseconds
13	The _____ was developed by John Presper Eckert and John W._____ in the year 1946	ENIAC Mauchly

User defined patterns are generated to extract the two consecutive “noun” tag from the sentence to frame the questions. As well generate the pattern like, “characteristics of”, “features of”, “benefits of”, “uses of”, “purpose of” to generate the subjective type questions based on blooms taxonomy. Then search whether the pattern appears in the document and once it matches the respective questions will be generated.

2) Dependency parser: Dependency parser is the term used to determine the dependencies between the terms of a phrase in order to assess its grammar and syntax [28]. As a result, the sentence is broken into constituent parts. The process is predicated on the premise that each linguistic unit in a phrase is inextricably linked to the next. These connections are referred to as dependencies. The relationships between each linguistic unit, or word, in a phrase are represented by directed arcs in a structured dependency. A dependence tag indicates the relation between different words.

In the proposed system, dependency parsing allows us to construct a parsing tree using tags to determine the relationship between words in a sentence rather than using any grammar rule as in syntactic parsing, which provides a lot of flexibility even when the order of words changes. These tags are used to identify the subject, verb, adverb, object in the sentence. The various rules have been generated to establish automatic objective type ‘wh’ questions.

The simple sentence has been parsed using dependency parser identification. The relation between the word exists in the sentence are identified through the tags like nsubj, vbd, jj, obj and amod as given in the dependency parser Fig. 2. Rules were created to generate short answer assessment questions.

By applying NER to the input sentence, “Rome won famous medal” where “Rome” is identified as “Location”. and the possibly generated questions will be “When did Rome won?”, but the generated question is not a semantically exact. In existing system, many researchers have developed the assessment questions based on NER for the short answers, but always cannot expected of holding location, person, cardinal number in the corpus. Hence in the proposed system, NER, Dependency parser and spacy matcherput together to produce the assessment questions. The proposed system produces possible number of questions by following the generated rules.

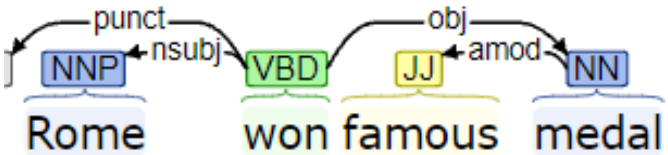


Fig. 2. Dependency Parser of Sample Sentence.

3) Rules based on Dependency Parser: The tag like nsubj, vbd, jj, obj, amod and det are determined to frame the rules as mentioned below.

Rule 1:

On the basis of the subject, the question has been generated as given. Use the Part-of-Speech and dependency parser to identify the tag of each word in the sentence. When the subject tag is the "person" then the keyword "Who" is added with the identified root tag and direct object(dobj) of the sentence, ended with question mark.

"Who" +root+ dobj

"Who "+root+ prep+ det+ pobj

Input sentence: Rome won famous medal

Possibly generated questions are,

Output Question: Who won medal?

Rule 2:

The question has been generated based on the Noun (NN) tag. The Noun and object tag of the sentence is determined by part-of-speech and dependency parser. Hence the keyword "What did" or "What" has been added with the subject and root tag of the given sentence.

"what did "+sub+root.

"what" + sub+ root.

As per the given input mentioned in Rule 1, possibly generated questions were,

Output: what did Rome won?

Rule 3:

Open clausal complement (xcomp), a verb or an adjectiveis a predicative or clausal complement that does not have its own subject. It also involved in question generation using the keyword "who" added with the auxiliary, root and xcomp tag of the components from the sentence.

"Who "+aux+root+"ing to "+xcomp.

The input sentence: "The Sheriff try to eat the apples while the outlaws were fasting", as per the rule 3, the generated question is

Output: Who were trying to eat?

Rule 4:

When any place or location has mentioned in the sentence then the keyword "Where did" or "Where" has added with the subject and root of the sentence to generate the questions.

"Where did "+sub+root.

"Where" + sub+ root.

Rule 5:

Check whether the word is the abbreviated. All the abbreviated words are in uppercase, if so then generate question to give the expansion of the abbreviations.

"Write the expansion of" + abbreviated word

Write the expansion of ENIAC?

Rule 6:

Apply the match pattern to extract the continuous two noun words, so as to generate the semantically correct assessment question. Usually, the single word of the sentence will be extracted either using NER/POS/Dependency Parser. Sometime the two sequential word gives the completeness. Example "data structure", using NER/POS/Dependency parser the term "data" and "structure" tokenized separately. Hence the matcher is defined with NLP,

pattern = [{"POS": "NOUN"}, {"POS": "NOUN"}], so that two continuous "nouns" have been extracted to generate the questions.

Rule 7:

Create a template like "uses of ", "examples of", "features of", "characteristics of", "types of", "methods of". Check if the input corpus consists of any of this template, when it holds then use the blooms verbal list to frame subjective type assessment questions. The sample input corpus is given below,

Input: "Array is a kind of data structure store a fixed-size sequential collection of elements of the same type. Array has the characteristics of storing the contiguous element".

The POS of the given is input determined, Array is termed as "NN" as given in Fig. 3, the predefined word combined with NN to frame assessment questions.

The dependency parsers is identified for the same input corpus as given in Fig. 4, and extract the subject of the sentence. Not all the noun(NN) will be fine to involve for questions, so the noun has to be the subject of the sentence and the semantically meaningful questions can be generated as per NER and dependency parser.

Array is the determined as Noun (NN) tag and it is the subject (nsubj) of the input sentence.

Affording to the input corpus, "characteristics of" is present which can be combined with the noun tag to frame assessment questions.

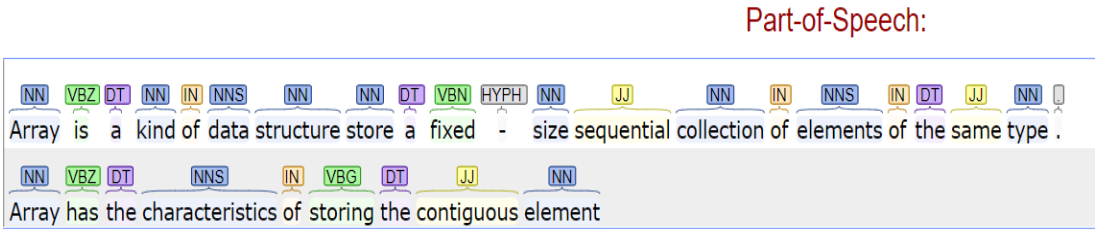


Fig. 3. Part-of-Speech.

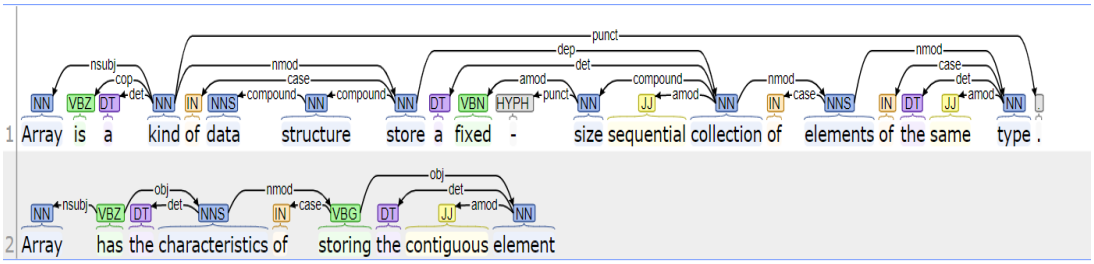


Fig. 4. Dependency Parser.

Blooms verbal list like, “List”, “write”,” State”, “Explain”, “describe”, “Discuss”, “Illustrate”, “Elaborate”, “Mention”.

Rule: Blooms verbal + “the” + ” predefined text” + NN.

Output:

1) Mention the characteristics of Array.
2) State the characteristics of Array
3) List the characteristics of Array

The different set of rules was implemented to the sentence of the input corpus to generate the possible number of assessment questions.

H. Experimental Result

Proposed system has been executed for the sample corpus taken from Grade V computer science book with 50 different text corpuses. Each chapter consists of topics and sub-topics. The system generated a range of objective and subjective questions of various categories. The generated questions have been manually crisscrossed by domain experts, and the accuracy (A), Recall (R), and Validity (V) of the questions have been measured.

Accuracy (A) is defined as the proportion of retrieved examples that are relevant. It is a metric that measures the number of correct positive predictions that are made during the execution.

$A= Q_v/(Q_v+ Q_i)$

Recall (R) is a metric that quantifies the proportion of correctly predicted positive predictions out of all possible positive predictions.

$R=Q_v/(Q_v+ Q_{ng})$

Validity is a metric that represent the number of valid and meaningful questions generated by the proposed system have been compared with the text book questions. Cardinality and validity of each question type is given in Table VIII. Table IX shows the accuracy and recall of the corpus.

TABLE VIII. CARDINALITY AND VALIDITY OF PROPOSED SYSTEM

Corpus Name	Fill up the double blank space		True or False		Short Question Answer	
	Cardinality	Validity	Cardinality	Validity	Cardinality	Validity
ENIAC	12	13	51	48	10	10
EDSAC	13	15	48	45	12	13
EDVAC	11	13	52	50	10	9
UNIVAC	14	15	55	54	11	09
Microcomputers	10	12	46	44	7	8
Minicomputers	13	14	42	43	8	6
Mainframes	16	17	45	41	9	8
Supercomputer	15	17	47	45	10	9
Optical Character Reader	12	13	43	42	9	7
Optical mark Reader	13	14	45	43	8	7
Cardinality	**129**	**143**	**474**	**455**	**94**	**86**

TABLE IX. ACCURACY AND RECALL OF PROPOSED SYSTEM

Corpus Name	Fill up the double blank space		True or False		Short Question Answer	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
ENIAC	0.92	0.86	0.96	0.88	0.88	0.88
EDSAC	0.87	0.93	0.94	0.94	0.93	0.82
EDVAC	0.85	0.85	0.96	0.96	1.0	0.90
UNIVAC	0.93	0.93	0.98	0.95	0.81	0.75
Microcomputers	0.83	0.83	0.96	0.96	0.88	0.80
Minicomputers	0.93	0.87	1.02	0.98	0.83	0.71
Mainframes	0.94	0.94	0.91	0.93	0.86	0.81
Supercomputer	0.88	0.94	0.96	0.96	0.80	0.80
Optical Character Reader	0.92	0.92	0.98	0.95	1.0	0.85
Optical mark Reader	0.93	0.87	0.96	0.93	1.0	0.92
	9.01	**8.94**	**9.63**	**9.44**	**9.06**	**8.24**

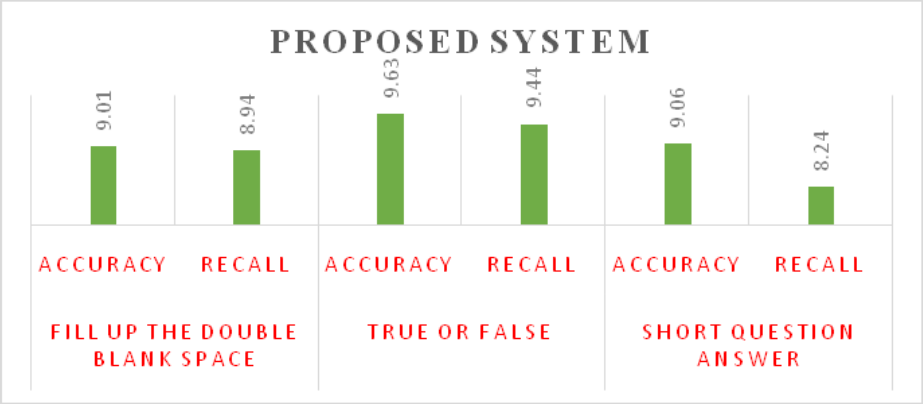


Fig. 5. Accuracy and Precision.

Accuracy and recall of the proposed ssytem is given in Fig. 5. The accuracy of each type is nearly 0.9 % and recall is nearly 0.8%. The overall accuracy and recall are 0.92 % and 0.88 %. The proposed system has been compared with existing system [29][30] and comparitvely produces good result. By the existing system[30], the overall precision and recall for the subjective type question were 0.85% and 0.8 %.

IV. CONCLUSION

The proposed system has been designed to verify the learner's knowledge. Its main purpose is to generate objective and subjective type questions from unstructured text. Better result has been achieved with accuracy level of 0.9 and recall as 0.8. Subject expert, recommended the system to assess the learner of grade V student in practice. In future the system can be elaborated with more rules to generate possible number of questions. In match the following question, the important word and their description can be considered to frame questions.

REFERENCES

[1] Ali, Y. Chali, and S. A. Hasan. Automation of question generation from sentences. 2011.

[2] Beg and A. Beg, Using open technologies for automatically creating question-and-answer sets for engineering MOOCs, Comput. Appl. Eng. Educ. 26 (2018), 617–625.

[3] L. Bednarik and L. Kovacs, Implementation and assessment of the automatic question generation module, 3rd International IEEE Conference on Cognitive Infocommunications (CogInfoCom), Inst. Electr. Electron. Eng., Kosice, Slovakia, 2012, pp. 687–690.

[4] Ming Liu, A. Calvo Rafael, and V. Rus, Automatic question generation for literature review writing support, International Conference on Intelligent Tutoring Systems, Springer, 2010, pp. 45–54.

[5] P. Pabitha, M. Mohana, S. Suganthi, and B. Sivanandhini, Automatic question generation system, International Conference on Recent Trends in Information Technology (ICRTIT), Inst. Electr. Electron. Eng., Chennai, India, 2014, pp. 1–5.

[6] K. Antonis, T. Daradoumis, S. Papadakis, and C. Simos, Evaluation of the effectiveness of a web-based learning design for adult computer science courses, IEEE Trans. Educ. 54 (2011), 374–380.

[7] J. C. Brown, G. A. Frishkoff, and M. Eskenazi, Automatic Question Generation for Vocabulary Assessment, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (Vancouver, British Columbia, Canada, 2005), Assoc. Comput. Linguist., Stroudsburg, PA, USA, 2005, pp. 819–826.

[8] Maor, and J. K. Currie, The use of technology in postgraduate supervision pedagogy in two Australian universities, Intl. J. Educ. Technol. Higher Educ. 14 (2017), 1–12.

[9] Kurtasov, A system for generating cloze test items from Russian-language text, Proceedings of the Student Research Workshop associated with RANLP 2013, INCOMA Ltd., Shoumen, Bulgaria, Hissar, Bulgaria, pp. 107–112.

[10] J. Lee and S. Seneff, Automatic generation of cloze items for prepositions, Proceedings of Interspeech, International Speech

Communication Association (ISCA), Antwerp, Belgium, 2007, pp. 2173–2176.

[11] M. Yildirim, A genetic algorithm for generating test from a question bank, Comput. Appl. Eng. Educ. 18 (2010), 298–305.

[12] Mannem Prashanth, Rashmi Prasad, and Aravind Joshi. (2010). Question generation from paragraphs at UPenn: QGSTEC system description. In Proceedings of QG2010: The Third Workshop on Question Generation, pages 84–91.

[13] Susanti, Y., Iida, R., Tokunaga, T. (2015). Automatic generation of English vocabulary tests. In Proceedings of the 6th International Conference on Computer Supported Education (CSEDU 2015), pp.77-87, 2015.

[14] Colleen E Crangle and Joyce Brothers Kart. (2015). A questions-based investigation of consumer mental Health information. PeerJ, 3:e867.

[15] Zhe Liu and B. J. Jansen, Subjective versus objective questions: Perception of question subjectivity in social Q&A, International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2015, pp. 131– 140.

[16] M. Divate, and A. Salgaonkar, Automatic question generation approaches and evaluation techniques, Curr. Sci. 00113891 (2017), 113.

[17] A. Narendra, M. Agarwal, and R. Shah, Automatic Cloze- Questions Generation, Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria 2013, pp. 511–515.

[18] M. Majumder, and S. K. Saha, Automatic selection of informative sentences: The sentences that can generate multiple choice questions, Knowl. Manage. E-Learn. Intl. J. 6 (2014), no. 4, 377–391.

[19] Swart (2010). Evaluation of Final Examination Papers in Engineering: A Case Study Using Bloom's Taxonomy. IEEE Transactions on Education, (May 2010) Vol. 53, No.2 257-264.

[20] G.Deena and K.Raja , "Sentence Selection Using Latent Semantic Analysis for Automatic Question Generation in E-Learning System", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019.

[21] Wolfe, J.H.: Automatic question generation from text-an aid to independent study. ACM SIGCSE Bull. 8(1), 104–112 (1976).

[22] Sheetal Rakangor et al," Computer Aided Environment for Drawing (To Set) True or False Objective Questions From Given Paragraph", Int. Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 4, Issue 5(Version 3), May 2014, pp.105-108.

[23] Michael Heilman and Noah, A Smith. 2009. Question generation via over generating transformations and ranking. Technical Report CMU-LTI-09-013, Language Technologies Institute, Carnegie Mellon University.

[24] Smith, S. ,Avinesh, P.V.S. and Kilgarriff, Adam (2010) 'Gap-fill Tests for Language Learners: Corpus-Driven Item Generation' Proceedings of ICON-2010: 8th International Conference on Natural Language Processing. held: 8-11 December 2010, Kharagpur, India. India: Macmillan Publishers.

[25] NikiforosKaramanis, Le An Ha and Ruslan Mitkov. 2006 Generating Multiple-Choice Test Iterms from Medical Text: A Pilot Study, In Proceedings of INLG 2006, Sydney, Australia.

[26] Ruslan Mitkov, Le An Ha and NikiforosKaramanis. 2006 A computer-aided environment for generating multiple-choice test items, Natural Language Engineering 12(2): 177-194.

[27] G.Deena and K.Raja , "A Novel Dynamic Self-Assessment Question Preparation Mechanism Using Natural Language Processing", Two Day International Virtual Conference on Contemporary Practices of Technology and Management for Economic Growth (ICTMEG'20).

[28] M.C.DeMarneffe and C. D. Manning, Stanford typed dependencies manual. Technical report, StanfordUniversity, 2008.

[29] Shivank Pandey and KC Rajeswari. Automatic question generation using software agents for technical institutions. International Journal of Advanced Computer Research, 3(13):307–311, 2013.

[30] ShivaliJoshi, Parin Shah , Sahil Shah, "Automatic Question Paper Generation, according to Bloom's Taxonomy, by generating questions from text using Natural Language Processing", International Journal of Innovative Science and Research Technology, Volume 6, Issue 4, April – 2021.

IoT based Date Palm Water Management System Using Case-Based Reasoning and Linear Regression for Trend Analysis

Ferddie Quiroz Canlas, Moayad Al Falahi, Sarachandran Nair
Muscat College, Oman

***Abstract*—Palms trees (Phoenix dactylifera L.), Al Nakheel in Arabic are known to have cultural and economic importance to Gulf and Arabic-speaking countries. However, using the traditional method of cultivation, improper use, and depletion of water is perceived as the major challenge as farmers used almost two and a half times the required amount without considering numerous factors. This paper attempts to develop an implementation model of a water management system for Date Palm Trees using Case Based-Reasoning. The said model involves an IoT-based module comprised of NodeMCU, soil moisture, temperature, and humidity sensors that automate the settings of the water amount for the whole year based on palm age, temperature, air humidity, and soil moisture. CBR calculates the amount of water supplied to palm trees (based on initial knowledgebase cited from previous empirical studies) and stores it in a cloud-based database. These data and hardware status can be accessed using a mobile application. When the temperature or soil moisture sensor fails, data trends are retrieved from the database and processed using Linear Regression Analysis. The test results have shown that the proposed model helped in a significant decrease in water consumption compared to the traditional method.**

Keywords*—Date palm tree; case-based reasoning; IoT; mobile application; NodeMCU; water management system

I. Introduction

Palms trees (Phoenix dactylifera L.) also known as Al Nakheel in Arabic are considered as one of the most famous plantations in Gulf countries, such as Oman and Saudi Arabia, and some Arabic-speaking countries in Africa such as Egypt [1]. There is also a prevalence of the tree outside the Arab countries such as Spain, Australia, and the USA [2].

Besides oil and gas, date palm products are one of the income sources of Oman. Date palm is the primary agricultural crop in the country, and it constitutes 80% of all fruit crops produced and 50% of the total agricultural area in the sultanate. Oman is the eighth largest producer of dates in the world with an average annual production of 260,000 metric tons [3].

As part of Oman's 2040 vision and national priorities, the government would like to diversify its sources of income and not rely alone on fossil fuels [4]. To contribute to this vision, the Ministry of Agriculture and Fisheries spearheaded the "One Million Date Palm Trees Project" aiming to revitalize the agriculture sector to enable the country to achieve food security and to drive the economy [5].

However, in achieving this, the Ministry faces many challenges. According to Al Marshudi [6] as cited in Al Yahyai et al. [3] Date production in Oman is still traditional from irrigation to the methods of applying fertilizers. Due to a subtropical dry, hot desert climate with low annual rainfall and very high temperatures in summer, the insufficiency of quality and quantity of water is the major concern not only of Dates farmers but the rest of the agricultural sector [3]. Water management in farming is a major challenge according to the Middle East Desalination Research Centre (MEDRC) [7].

This paper attempts to develop an implementation model of a water management system for Date Palm Trees using Case Based-Reasoning. The said model involves an IoT-based module that automates the settings of the water amount for the whole year based on palm age, temperature, air humidity, and soil moisture. CBR calculates the amount of water supplied to palm trees and then stores it in a cloud-based database. When the temperature and soil moisture sensors fail, data trends are retrieved from the database and processed using Linear Regression Analysis.

II. Review of Related Literature

There is a wide array of past and current studies on water management in various contexts and parameters. Recent papers published are categorized either into IoT-based with no artificial intelligence approaches involved or the combination of both (with emphasis on fuzzy logic). This paper also cited the latest studies on the application of case-based reasoning to irrigation systems.

Qomaruddin et al [8] proposed a watering system for greenhouses involving air temperature, air humidity, and soil moisture as input parameters. For the user to access the system and control the supply of water remotely, the MQTT (Message Queue Telemetry Transport) protocol bridges the user's mobile phone and the Wemos D1 Uno microcontroller. The microcontroller processes inputs from the sensors and feeds the data to AdaFruit, a third-party IoT webpage that users access using their smartphones. Since there is neither algorithm nor complete automation involved, the user has to perform decisions based on the data displayed on the website and water the plants accordingly. A similar study utilizing the AdaFruit webpage for IoT aimed to predict plantation and crops' health and then notifies farmers through emails. The Arduino, connected to WIFI using third-party hardware, collects data from soil moisture, pH, flame, and humidity sensors. These data are then processed based on a rule-base using the FindS

algorithm to predict the plants' health. Based on the health status, the farmer decides the right amount of water supplied to plants [9]. Similarly, the apparent drawback of the study is the necessity for the farmer to regularly check emails and be physically present to water the plants.

Sweety et al. [10] suggested the use of a rule base in control of the watering process in gardens. Their prototype involved a PIC microcontroller-based module interfacing temperature, humidity, and soil moisture sensors that enable the users to control the motor via Bluetooth. The motor pump turns on when the temperature level is between 35 and 40 degrees centigrade, humidity is more than 35%, and soil moisture is above 100MA. There is neither clear basis nor references cited on the threshold values used, and the user must be within the vicinity to perform the control of the system since Bluetooth has a limited distance covered.

A group of researchers [11] from the Universitas Klabat in Indonesia found out that automating water supply to plants maintained the soil moisture at an average of 62%. Their prototype consists of soil moisture sensors, Wemos D1 Microcontroller WiFi enabled, Relay, and solenoid valve. Watering starts when moisture is detected to be between 30-35% and then stops afterward. The Blynk Apps, a third-party IoT platform, monitors the status of the microcontroller and feeds data to ThingSpeak, another IoT Platform for storage. Users must log in to these platforms to view data. The study suggested the development of a dedicated software application as part of future work as it relies heavily on third-party platforms. There are no clear bases to support the claim that 62% moisture is the ideal level for sustaining plants' water needs. Also, the study admitted the absence of temperature and humidity sensors as a limitation. Likewise, another project [12] employed ThingSpeak functioning as a cloud server to record all the data and link the hardware prototype with the android application to irrigate plants, flowers, and crops. Using three microcontroller modules (Arduino UNO, NRF24L01, NodeMCU) communicating with ThingSpeak through WIFI, the user can monitor the performance of the system. The said microcontrollers turn on the pump when sensors detect 20% moisture until it reaches 71% level. The study was successful in achieving and maintaining the desired soil moisture level. However, no pieces of literature support the indicated maximum moisture level as the ideal threshold value. Wahid et al., [13] and Ahmmad et al., [14] proposed the same project using NodeMCU Lua and soil moisture, humidity, temperature, and light sensors.

Ying and his colleagues [15] from the Universiti Tun Hussein Onn Malaysia developed a web-based water management system for oil palm nurseries written in PHP, enabling users to obtain reports on the right amount of water required by palm trees. The user inputs the parameter values for the rainfall and watering time, analyzed using Fuzzy Logic. The MySQL database stores the data for later retrieval. The project has no hardware implementation. Also, there was no indicated basis for the fuzzy rules. Similarly, a MatLab simulation of a greenhouse irrigation system involving Mamdani Fuzzy Logic and temperature, humidity, wind speed, and radiation influence as inputs helped to achieve efficient energy consumption, desired soil moisture level compared to an ON/OFF controllers, and a cost-efficient prototype [16]. On the other hand, university research collaboration in Kenya found out that the Sugeno inference system is better than the Mamdani in analyzing relative humidity, temperature, sunshine illumination, solar radiation, and wind speed to infer the desired amount of water. The MatLab simulation determines the duration of the revolution of the motor pump leading to more efficient water and energy consumption as compared to an ON/OFF system [17]. Similarly, Oubehar et al., [18] proposed a Matlab-based intelligent control for greenhouse using ANFIS technology.

A recent study by Zhai et al. [19] proposed a case-based reasoning model accurately predicting the amount of water for Grape farming. The model involved various parameters such as solar radiation, temperature, humidity, wind speed, rainfall, precipitation, soil moisture, and grape growth stage. Moreover, the said model is ideal for hardware and software implementation. Another similar study by Krongtripop et al. [20] proved that CBR-based water control systems implemented in mixed gardens are better than the traditional time-based approach in terms of water conservation and the height of the trees. The prototype involved temperature and soil moisture as parameters to be fed into the Arduino Uno R3 Microcontroller connected to WIFI using third-party hardware Xbee Wireless Module. Data are collected and stored on a computer server. A desktop application written in C# controls the system and queries the database for the CBR process. The perceived drawback of the study is the need for the server to be in the same proximity as the hardware module, hence requiring physical human intervention.

There are few recent studies on water management using case-based reasoning, and mostly they are introducing the Mathematical model for adoption. Fuzzy Logic is a commonly used algorithm for predicting values based on some input parameters. However, despite its popularity, there are known drawbacks to this algorithm. A critical review of Behrooz et al. [21] enumerated the following disadvantages: (1) stability is not assured (2) designers must perform a series of trials and errors to achieve optimal output and (3) the presence of several tuning parameters. Moreover, in Fuzzy Logic, all parameters are treated with equal weights, and the process is confined to a particular use case, thus less dynamic as compared to other algorithms [22].

III. Methodology

A. Requirement Analysis and Locale

The Ministry of Agriculture, Fisheries, and Water Resource (MoAFWR) of the Sultanate of Oman is the agency having jurisdiction over the cultivation of Date palm trees. The proponents collaborated with an expert from the MoAFWR about the traits, water needs, and other parameters (which were not covered in this study) involved in growing the said palm trees. The Ministry also confirmed that no similar project had been proposed or currently used in the Sultanate. Also, the same expert performed the testing of the proposed prototype.

B. Proposed Architecture

The proposed architecture, as shown in Fig. 1, consists of seven major components. The NodeMCU that has an on-board

WIFI module collects information about humidity, temperature, and soil moisture levels, which are then processed by a web-based Case-Based Reasoning (CBR) system written in PHP. The Apache Web Server housed these PHP scripts that transact with the MySQL Database Server containing all the sensor reading and water levels supplied based on the CBR process.

Fig. 1. Architecture of the Proposed Model.

The mobile application (running within the Android operating system) developed using Android studio enables the user to monitor the hardware status (see Fig. 2), sensor readings, and water level trends (parsed in a JSON array form) from the MySQL database. Similarly, the user can monitor the status of the hardware components (e.g., the microcontroller and the sensors), which the Firebase online database stores in real-time.

Fig. 2. Mobile Application Interface for Hardware Components Monitoring.

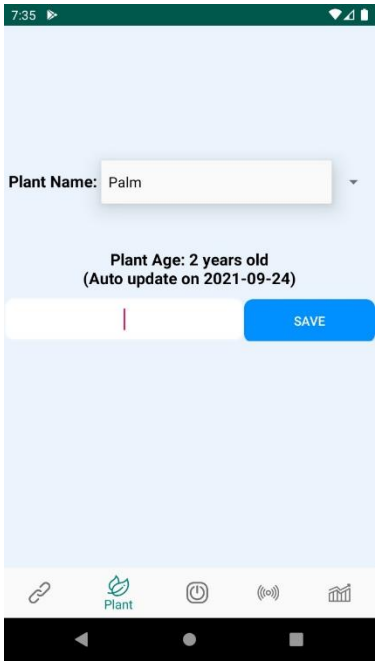


Fig. 3. Mobile Application Interface for user Inputs.

Within the same application, users can input Date palm tree details such as age and watering frequency. Fig. 3 shows the mobile application interface for user inputs.

For testing purposes, the prototype used free web-based tools and database services.

C. Major Hardware and Software Components

1) The hardware schematic diagram: The hardware prototype has six (6) major components. Fig. 4 shows all the components numbered and whose descriptions are found in Table I.

The prototype, as much as possible, is consists of generic components with specifications listed in Table I.

All hardware components are interconnected using a breadboard with exceptions to some that require jumper wires. There are two power sources in the prototype, as indicated in Fig. 4. The USB Micro-B Cable powers up the microcontroller and the sensors. Since the voltage coming from the output pin of the microcontroller cannot directly supply the submersible water pump, the relay module routes the power from the external 12 volts power supply to the pump.

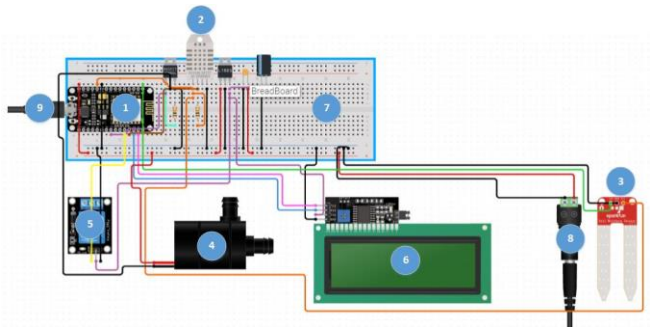


Fig. 4. Major Hardware Components.

TABLE I. MAJOR HARDWARE COMPONENTS

Hardware Name	Description
(1) NodeMCU	Arduino compatible ESP8266 based microcontroller with on-board WiFi
(2) DHT22/11 Humidity and Temperature Sensor	Range: -40 °C to 125 °C, Humidity Range: 0% to 100%
(3) Soil Moisture Sensor	Returned Values: from 0 (completely dry) to 1023 (completely moist).
(4) Submersible Pool Water Pump	Rated voltage: 12V DC. Max rated current: 350mA. Max flow rate: 240L/H.
(5) Relay Module	Operating voltage: 5-12V
(6) LED Display (Serial Monitor)	LCD Display 20x4 I2C
(7) Breadboard	Full size solderless breadboard. It has 2 split power buses, 10 columns, and 63 rows.
(8) Power Supply (12V)	Water Pump power supply

At the initial stage, the WIFI SSID and key are manually embedded into the code to enable the microcontroller to connect to remote databases in real-time. The NodeMCU collects information from the temperature/humidity and soil moisture sensors and sends it to the webserver for CBR processing (or Linear Regression calculation in case one of the sensors fails) and in the MySQL database for storing through the HTTP POST protocol. Fig. 5 shows the partial code.

The web server returns CBR results to the NodeMCU, which then sends a signal to the output port that triggers the relay module to activate the submersible water pump.

D. Input / Output Parameters, Threshold Values and Database Design

The CBR calculation involves three inputs: temperature, humidity, and soil moisture. For brevity, Fig. 6 shows a partial NodeMCU code for reading from sensors, in this example is temperature.

```
//Compute water amount based on the CBR or MLR technique
float ComputeWaterAmount(String technique)
{
  http.begin(ROOT_URL);
  http.addHeader("Content-Type","application/x-www-form-urlencoded");

  String data = "temp_level="+String(temperature) +"&soil_mois_level="
  +String(soilMoisture)
  +"&plant_id="+String(plantID) +"&"+String(technique) +"=1"
  +"&hostname="+String(HOST_NAME) +"&username="+String(USERNAME)
  +"&password="+String(PASSWORD) +"&dbname="+String(DB_NAME);

  http.POST(data);
  String payload = http.getString();
  http.end();

  size = payload.length();
  char json[size+1];
  strcpy(json, payload.c_str());
  StaticJsonDocument<20000> doc;
  DeserializationError error = deserializeJson(doc, json);

 if(error)
  {
    Serial.println("-------------------------------------------------");
    Serial.print(F("deserializeJson() failed: "));
    Serial.println(error.c_str());
    Serial.print("Mysql connection error!");
    Serial.println(payload);
    return -1;
  }

  //update device status
  Firebase.setBool(firebaseData, "Records Updated", true);
  return doc["water_amount"];
  doc.clear();
}
```

Fig. 5. NodeMCU Code for CBR / LR Processing and Storing Data to MySQL.

```
//Get temperature value
float getTemperature()
{
  Serial.println("\n\n*******************************************************");
  Serial.print("Temperature: ");
  str = "0";
  startMillisTimer = millis();
  while(true)
  {
    Firebase.getString(firebaseData,"System Status",systemStatus);
    Firebase.getBool(firebaseData,"Connection Request",connectionRequest);
    Firebase.getBool(firebaseData,"Disconnection Request",disconnectionRequest);
    if(systemStatus=="Off" || connectionRequest || disconnectionRequest)
      break;
    if(Serial.available())
    {
      str = Serial.readString();
      Firebase.setBool(firebaseData, "Temperature Sensor Status", true);
      tempSensorRunning = true;
      break;
    }
    millisTimer = 10UL * 1000UL;
    if(millis() - startMillisTimer >= millisTimer)
    {
      Firebase.setBool(firebaseData, "Temperature Sensor Status", false);
      sendMessage("Temperature sensor is not available/working");
      tempSensorRunning = false;
        break;
    }
  }
  value = str.toFloat();
  return value;
}
```

Fig. 6. NodeMCU code for Temperature Reading.

CBR takes the temperature values supplied by the sensor directly. However, for better presentation, the soil moisture value is normalized using (1) and converted to a percentage form.

$$Soil\ Moisture\ (\%) = \frac{Actual\ Value}{1023} * 100 \quad (1)$$

The CBR calculation returns the water amount in liters. In this situation, there is a need to convert liter to revolution time. The water pump, as indicated in Table I, supplies 240 liters of water in an hour. Thus, to obtain revolution time:

$$Revolution\ Time\ (H) = \frac{Amount\ of\ Water\ (L)}{240} \quad (2)$$

The Center for Study of the Built Environment (CSBE) [23] in Jordan reported that young Date palm trees consume 20 to 25 liters, and the mature ones consume around 40 liters of water, respectively. These amounts of water enable palm trees to yield more fruits (in addition to soil fertility) and become healthy. The frequency of supplying water to Date palm trees varies depending on the age of the tree. Table II shows the findings of Al Hyari [24].

The baseline or initial knowledge base for CBR, to be used for the Retrieval Process (Global and Local Similarities Calculations), were taken from the findings of Bhat et. al. [25] based on the Penman-Monteith Equation. Daily prescribed water consumption and calculated daily evapotranspiration are combined to obtain an adequate amount of water to palm trees and to avoid loss due to evaporation. Table III shows the newly calculated data with other relevant parameters and soil moisture is set to 0%.

TABLE II. WATERING FREQUENCY CONDITIONS

Temperature	Age	Watering Frequency & Amount
High	>4 years	1 time/week, more water
Low	>4 years	1 time/week, less water
High	≤4 years	Every 3 days, more water
Low	≤4 years	Every 3 days, less water

TABLE III. SUGGESTED WATER QUANTITY

Temperature (°C)	*Humidity (%)*	*ET Coeff (%)*	*Water Amount (L)*
23.4	54	4	37.27
23.6	60	3	37.25
27.5	49	5	37.81
37.8	20	9	39.26
46.4	16	10	39.53
48.3	11	14	40.95
50.6	14	14	40.94
44	16	11	40.07
35.2	6.89	7	38.68
32	48	5	37.79
24.8	42	4	37.57

E. Application of Case-Based Reasoning

Case-Based Reasoning (CBR) is a machine learning algorithm [26] that manifests human expertise and can work effectively with criteria-based comparison [27] of the new cases and the existing solved problems stored in the knowledge base [28]. CBR, as shown in Fig. 7, has four major phases: Retrieve, Reuse, Revise, and Retain. These phases are done in an iterative fashion adding new solved problems to the knowledge base, thus making it more intelligent [29].

The retrieval phase is the center of CBR. Existing cases retrieved from the knowledge base are compared to the new problem using Local Similarity (LS) and Global Similarity (GS) calculations. LS comprised is used to break down the problems' attributes and compare them to existing ones in the database. These attributes can be Discrete (3) or Continuous (4) in nature and take different calculation approaches [30].

Discrete Values

$$sim\,(a,b) = \begin{cases} 1\ if\ a = b \\ 0 \\ 0\ if\ a \neq b \end{cases} \quad (3)$$

where,

a is the new feature, and

b is the previous feature

Continuous Values

$$sim(a,b) = 1 - \frac{|a-b|}{range} \quad (4)$$

where,

a is the new feature,

b is the previous features, and *range* is the value of the difference between the upper and lower boundary of the set.

Global Similarity GS (5), on the other hand, takes the build-up of all the local similarities and is used to make a generalized comparison of each problem. This paper expresses global similarities in percentages.

$$sim(A,B) = \frac{1}{\sum w_i} \cdot \sum_{i=1}^{p} w_i \,.\, sim_i(a,b) \quad (5)$$

where,

A is the new case,

B is the previous case,

a is the new feature from the local similarity

Once CBR found an ideal solution in the knowledge base, it proceeds to the Reuse stage to adopt it, or to revise, wherein users formulate the necessary solution to the new problem. This new problem and solution are added to the knowledge base during the Retain phase. Fig. 8 shows the CBR implementation in PHP.

F. Linear Regression Analysis

The study used Linear Regression Analysis as a backup mechanism in a situation wherein one of the sensors fails. Setting one of the input parameters to a null value, CBR exhibits faulty behavior affecting the integrity of the knowledge base. The results of the Correlation analysis show that temperature has a "strong positive" relationship (r=0.799) with the water amount. Reversely, humidity and soil moisture have negative strong (r=-0.878) and negative, very strong (r=-0.951) correlations to water amount, respectively. With the recorded sensor reading and the amount of water supplied for the current month, Linear Regression Analysis predicts the right amount of water, using temperature as the predictor.

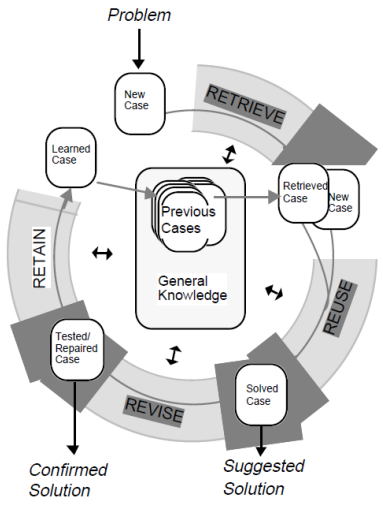


Fig. 7. The Case-Based Reasoning Cycle.

```
//local similarity calculation

function LocalSimilarityContinuous($input, $retrievedData, $range)
{
    $LS=1-abs(($input-$retrievedData))/$range;
    return $LS;
}

//global similarity
function GlobalSimilarity($tempLevel,$humidity,$soilMoisLevel)
{
    $GS=((1*$tempLevel)+(1*$humidity)+(1*$soilMoisLevel))/3;
    return $GS;
}
```

Fig. 8. CBR Implementation in PHP.

G. Database Design

There are three (3) database tables used in this study: (1) for hardware status information, (2) palm tree information, and (3) CBR knowledgebase. Fig. 9 shows the Firebase table structure that stores the real-time status of the hardware components such as the soil moisture, temperature, and humidity sensors. The data is displayed in the mobile application for monitoring purposes.

Fig. 9. Firebase Table for Hardware Status.

Fig. 10 shows the table holding the information for the plant. The plant age plays important role in determining the frequency of watering.

#	Name	Type	Collation
1	plant_id	int(11)	
2	plant_name	varchar(50)	utf8mb4_general_ci
3	plant_age	int(11)	
4	next_update_date	date	
5	watering_frequency	int(11)	

Fig. 10. MySQL Table for Palm Tree Details.

Fig. 11 shows the knowledge-base table where the actual reading of the soil moisture, temperature, and humidity sensors was recorded. The CBR algorithm depends on these data.

#	Name	Type	Collation
1	trend_id	int(11)	
2	plant_id	int(11)	
3	temp_level	float	
4	humidity	float	
5	soil_mois_level	float	
6	water_amount	float	
7	sensors_status	varchar(3)	utf8mb4_general_ci
8	date_time	datetime	

Fig. 11. MySQL Table for CBR Knowledgebase.

IV. Results and Discussion

A. Test Case

Table IV contains the test case dataset. In this test case, the FreeCBR tool expedites the process of the CBR calculation and presentation of results. For clarity, the mature palm tree group was used in the test case.

For demonstration purposes, Test Case 1 is processed with FreeCBR. The CBR calculation in Fig. 12 shows that Test Case 1 has a 42.25% similarity with Case No. 11 stored in the knowledgebase. Therefore, the water amount for Test Case one will be (6):

$$WtrAmt_{TestCase1} = WtrAmnt_{CaseNo11} * Hit\% \quad (6)$$

$$= 37.57 * 0.4225 = 15.87L$$

Table V shows the complete CBR similarity results and the calculated amount of water.

The succeeding figures show the CBR results (Fig. 13) and data trends (Fig. 14) in the mobile application.

B. Test Results

The conventional method of cultivating Date Palm trees follows the same way as the other Gulf and Arab countries. Farmers supply an average of 40 liters of water a day regardless of the soil moisture condition, as suggested by CSBE [23]. The joint research of the Environment Agency Abu Dhabi (EAD) and New Zealand [31] revealed that two and a half times the intended water amount is supplied to date palm trees using the traditional way. That is equivalent to 100 liters of water.

TABLE IV. Test Case Dataset

Test Case No.	*Temp(°C)*	*Humidity (%)*	*Soil Moisture (%)*	*Water Amount (L)*
1	24	42	55	?
2	35	18	32	?
3	21	45	57	?
4	15	56	60	?
5	18	54	59	?
6	45	15	30	?
7	42	10	29	?
8	51	8	35	?
9	49	11	33	?
10	22	44	60	?

Case No	Temp in C	Humidity %	Soil Moisture %	Water Amount	Hit %
11	24.8	42.0	0.0	37.57	42.24699876193907
3	27.5	49.0	0.0	37.81	41.564938703719015
1	23.4	54.0	0.0	37.27	41.211812418207785
10	32.0	48.0	0.0	37.79	40.536791849301 99
2	23.6	60.0	0.0	37.25	40.128043411447
9	35.2	27.0	0.0	38.68	38.18684107328211
4	37.8	20.0	0.0	39.26	35.778653793563 53
8	44.0	16.0	0.0	40.07	32.156519971201966
5	46.4	16.0	0.0	39.53	31.022870225749323
5	48.3	11.0	0.0	40.95	28.986966548101346
6	50.6	14.0	0.0	40.94	28.55155605742804
7	50.0	12.0	0.0	40.87	28.39796743680889

Fig. 12. CBR Results for Test Case 1.

TABLE V. CBR-CALCULATED WATER QUANTITY

Test Case No.	CBR Knowledgebase		Hit (%)	Calculated Water Amount
	Case No.	Water Amount		
1	11	37.57	42.25	15.87
2	4	39.26	76.01	29.84
3	12	15.87	93.74	14.88
4	14	14.88	86.00	12.80
5	15	12.8	95.43	12.22
6	13	29.84	87.07	25.98
7	17	25.98	96.15	24.98
8	17	25.98	89.05	23.14
9	19	23.14	95.78	22.16
10	14	14.88	67.06	9.98

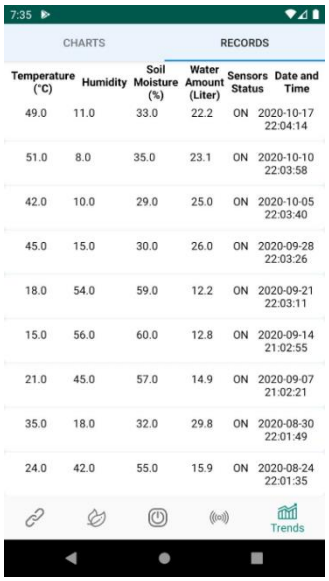


Fig. 13. Mobile Application Interface for CBR Results.

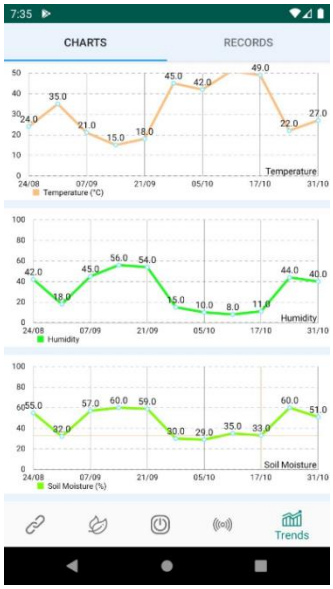


Fig. 14. Mobile Application Interface for Data Trends.

TABLE VI. WATER USAGE COMPARISON

Test Case No.	Water Amount (Liters)	
	CBR-Based	Conventional Approach
1	15.87	40
2	29.84	40
3	14.88	40
4	12.80	40
5	12.22	40
6	25.98	40
7	24.98	40
8	23.14	40
9	22.16	40
10	9.98	40
Total:	191.85	400

With the assumption of 40 liters of water usage in a single watering instance, Table VI shows the amount of water saved using the proposed model.

Using the CBR approach, water consumption significantly ($t=-9.621$, $p<0.05$) dropped down to 52% as compared to the traditional method.

V. CONCLUSION

Various literature indicates that date Palm trees have cultural and economic importance to Gulf and Arabic speaking countries. However, using the traditional method of cultivation, depletion of water is perceived as the major challenge.

Various studies attempted to propose prototypes and implementation models, but none of them provided semi-full automation and with lesser human intervention. Most of them provide software simulation and theoretical calculations without actual implementation. Although Fuzzy Logic is commonly used for automation similar to the study, it posed many challenges in the implementation. Also, there is a lacking basis to support threshold values for the rule-based algorithm used by previous studies to calculate the amount of water.

The proposed model had overcome the limitations found in previous studies by citing empirical studies identifying the various factors affecting effective date palm trees cultivation and using them as the initial knowledge base. The input parameters are processed using Case-Based Reasoning and Linear Regression for situations wherein one of the sensors fails to function. This backup feature is not present in the existing studies. The test results show a significant difference in water consumption, thereby helping to address the improper use of water resources.

The results of the study are confined to three parameters: soil moisture, temperature, and humidity. Other parameters such as sunlight intensity, rainfall, wind speed, etc., can be included. Also, the case-based reasoning table should be modified to accommodate other crops, hence making the system more dynamic.

REFERENCES

[1] L. I. El-Juhany, Degradation of Date Palm Trees and Date Production in Arab Countries: Causes and Potential Rehabilitation, Australian Journal of Basic and Applied Sciences, vol. 4, no. 8, pp. 3998-4010, 2010.

[2] A. Zaid and P. F. de Wet, Date Palm Cultivation, Food and Agricultural Organization of the United Nations, Rome, 2002.

[3] R. Al-Yahyai and M. Mumtaz Khan, Date Palm Status and Perspective in Oman, in Date Palm Genetic Resources and Utilization, Springer, 2015, pp. 207-240.

[4] B. K. Al Omari, Tourism Economics in Oman: A Statistical Study for the Period 2000-2017, Global Scientific Journals, vol. VII, no. 3, pp. 301-314, 2019.

[5] Oman Observer, Million Date Palm Trees Project to boost agriculture, Oman Observer, 18 March 2019. [Online]. Available: https://www.omanobserver.om/million-date-palm-trees-project-to-boost-agriculture/. [Accessed 10 August 2020].

[6] A. S. Al Marshudi, Oman traditional date palms: production and improvement of date palms in Oman, *Tropiculture*, vol. XX, no. 4, pp. 203-209, 2002.

[7] Times of Oman, Water management in farming challenge for Oman: MEDRC, Times of Oman, 25 February 2018. [Online]. Available: https://timesofoman.com/article/128923/oman/environment/water-management-in-farming-a-challenge-for-oman-medrc. [Accessed 10 August 2020].

[8] M. Qomaruddin, M. S. Mubarak and S. Mulyono, Plant Watering System on The Basis of Internet of Things (IoT) With Protocol of Message Queue Telemetry Transport (MQTT), in *The 18th International Conference on Positron Annihilation*, Orlando, 2019.

[9] A. Reghukumar and V. Vijayakumar, Smart Plant Watering System with Cloud Analysis and Plant Health Prediction, in *International Conference on Recent Trends in Advanced Computing 2019, ICRTAC 2019*, Chennai, 2019.

[10] J. Sweety, S. Dharshika, J. Jabez and M. Anu, An Enhanced Automation of Garden Watering Based On IOT, Global Journal of Pure and Applied Mathematics, vol. XIII, no. 6, pp. 2181-2191, 2017.

[11] J. M. Waworundeng, N. C. Suseno and R. R. Y. Manaha, Automatic Watering System for Plants with IoT Monitoring and Notification, Cogito Smart Journal, vol. IV, no. 2, pp. 316-326, 2018.

[12] F. Kamaruddin, N. N. N. Abd Malik, N. A. Murad, N. M. Abdul Latiff, S. K. S. Yusof and S. A. Hamzah, IoT-based intelligent irrigation management and monitoring system using arduino, TELKOMNIKA, vol. XVII, no. 5, pp. 2378-2388, 2019.

[13] N. Wahid and N. L. M. Azemi, Iot-Based Tropical Plant Monitoring, International Journal of Advanced Trends in Computer Science and Engineering, vol. IX, no. 1, pp. 395-399, 2020.

[14] S. N. Z. Ahmmad, M. M. Jaafar, F. Muchtar and M. I. Yusof, Automated Gardening Portable Plant Using IoT, International Journal of Advanced Trends in Computer Science and Engineering, vol. IX, no. 1, pp. 205-211, 2020.

[15] L. C. Ying, N. Arbaiy, M. Z. M. Salikon and H. A. Rahman, Plant Watering Management System Using Fuzzy Logic Approach in Oil Palm Nursery, Journal of Telecommunication, Electronic and Computer Engineering, vol. IX, no. 3, pp. 129-134, 2017.

[16] P. J. Kia, T. A. Far, M. Omid, R. Alimardani and L. Naderloo, Intelligent control based fuzzy logic for automation of greenhouse irrigation system and evaluation in relation to conventional systems, World Applied Science Journal, vol. VI, pp. 16-23, 2009.

[17] F. S. Ibrahim, D. Konditi and S. Musyoki, Smart Irrigation System Using a Fuzzy Logic Method, International Journal of Engineering Research and Technology, vol. XI, no. 9, pp. 1417-1436, 2018.

[18] H. Oubehar, M. El Khayat, I. Rkik, A. Ed-Dahhak, M. El Hassane Archidi and A. Lachhab, Intelligent control for an experimental greenhouse climate based on ANFIS Technology, International Journal of Advanced Trends in Computer Science and Engineering, vol. IX, no. 1, pp. 84-90, 2020.

[19] Z. Zhai, F. J. Martínez, N. L. Martínez and V. H. Díaz, Applying case-based reasoning and a learning-based adaptation strategy to irrigation scheduling in grape farming, Computers and Electronics in Agriculture, vol. CLXXVII, no. 105741, 2020.

[20] T. Krongtripop, P. Kirdpipat and W. Srigul, A Watering Controller System in Mixed Garden Based on Temperature and Moisture by Case Base Reasoning Technique via Wireless Network, RMUTI JOURNAL Science and Technology, vol. XI, no. 2, pp. 1-13, 2018.

[21] F. Behrooz, N. Mariun, M. H. Marhaban, M. A. Mohd Radzi and A. R. Ramli, Review of Control Techniques for HVAC Systems—Nonlinearity Approaches Based on Fuzzy Cognitive Maps, Energies, vol. XI, no. 495, pp. 1-41, 2018.

[22] M. S. Al Falahi, F. Q. Canlas and S. Nair, Water Control System for Al Nakheel Tree Using Fuzzy Logic with Trend Analysis, in 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Colchester, 2020.

[23] Center for Study of the Built Environment [CSBE], "Date Palm (Phoenix dactylifera)," csbe.org, [Online]. Available: 2020. [Accessed 27 September 2020].

[24] I. Al Hyari, Methods of Watering Palm Trees, Mawdoo, 14 May 2017. [Online]. Available: https://mawdoo3.com/%D8%B7%D8%B1%D9%82_%D8%B1%D9%8A_%D8%A7%D9%84%D9%86%D8%AE%D9%8A%D9%84. [Accessed 10 January 2020].

[25] N. R. Bhat, V. S. Lekha, M. K. Suleiman, B. Thomas, S. I. Ali, P. George and L. Al-Mulla, Estimation of Water Requirements for Young Date Palms Under Arid Climatic Conditions of Kuwait, World Journal of Agricultural Sciences, vol. VIII, no. 5, pp. 448-452, 2012.

[26] H. D. Burkhard, Case-Based Reasoning Technology from Foundations to Applications, Springer Verlag, 1998.

[27] F. Q. Canlas, Data Mining Model for Student Internship Placement Using Modified Case Based Reasoning, in Lecture Notes in Networks and Systems, vol. CXLIV, T. Masrour , I. El Hassani and A. Cherrafi, Eds., Springer, Cham, 2021, pp. 160-169.

[28] J. L. Kolodner, Case-Based Reasoning, San Francisco: Morgan Kaufmann, 1993.

[29] A. Aamodt and E. Plaza, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, AI Communications. IOS Press, vol. VII, no. 1, pp. 39-59, 1994.

[30] L. Chen, D. Ou and H. Yao, Research on Fault Diagnosis of Vehicle Equipment for High-speed Railway Based on Case-Based Reasoning, in 3rd International Conference on Electrical and Information Technologies for Rail Transportation (EIRT) 2017, Changsha, 2017.

[31] V. Todorova, Two and a half times more water used to irrigate palm trees than needed, thenational.ae, 31 July 2014. [Online]. Available: https://www.thenational.ae/uae/environment/two-and-a-half-times-more-water-used-to-irrigate-palm-trees-than-needed-1.468588. [Accessed 5 Septmeber 2020].I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

AquaStat: An Arduino-based Water Quality Monitoring Device for Fish Kill Prevention in Tilapia Aquaculture using Fuzzy Logic

Mark Rennel D. Molato
Institute of Information and Communication Technology
Isabela State University – San Mariano Campus, San Mariano, Isabela, Philippines, 3332

***Abstract*—In the Philippines, Tilapia fish farming sector is vital to the economy in providing substantial employment, income and meeting local demand for protein sources of the Filipinos. However, the possible benefits that can be derived from this industry are at stake because of the sudden occurrences of fish kill events. This can be attributed to a wide variety of natural and unnatural causes such as old age, starvation, body injury, stress, suffocation, water pollution, diseases, parasites, predation, toxic algae, severe weather, and other reasons. With the identified severe effects of fish kill events to the fish farmers, consumers and the fisheries industry, advanced measures and methods must be established to alleviate the adverse effects of this phenomenon. To solve the underlying problem on water quality monitoring system to improve freshwater aquaculture, various studies were already conducted. However, these studies merely focused on the reading and gathering of water parameters. In this paper, fuzzy logic was used to come up with a model that can analyze and generate result regarding the overall quality of the water being used in Tilapia aquaculture. The water parameters considered in this paper were temperature, dissolved oxygen, and pH level. The results of the water parameter readings using the conventional method were compared to the data that were gathered by AquaStat to test its accuracy and showed no significant difference. Also, the overall water quality obtained using the conventional method was compared to the overall water quality generated by AquaStat and obtained an accurate result.**

Keywords*—Fuzzy logic; fuzzy sets; fish kill; freshwater aquaculture; Tilapia

I. Introduction

The Philippines is positioned as the 8th top fish producing country in the world. It has a total inland resource of 749,917 ha as reported by the Bureau of Fisheries and Aquatic Resources employing a combined total of 1,614, 368 fishing operators nationwide. Moreover, aquaculture got the highest percentage of total fish production from 2018 – 2020 with 52.8% share to overall total fish produced [1]. This sector produces tons of fish, crustaceans, mollusk, and aquatic plants which contribute to the national economy of the country having the largest share next to agricultural crops at current and constant prices[2].

Tilapia (Oreochromis niloticus) ranked second among the fish species that are being produced in aquaculture in the Philippines [3]. Moreover, it is identified as the most consumed aquaculture fish in the Philippines for its taste and affordability [4], [5].

However, there is an indicative fall of the potential benefits of the aquaculture industry in the Philippines because of the occurrences of fish kill incidents directly harming fishing operators. The fish kill is a sudden mass death of the fish in any given body of water[6]. Also, it is defined as any unusual and noticeable increase in mortality due to infectious or non-infectious causes like oxygen depletion, pollutants, natural toxins and diseases in aquatic organisms [7]. It is characterized by large numbers of aquatic animals dying over a short time in any bodies of water where fish cultivation was being pursued [8].

Fish kills can be attributed to a wide variety of natural and unnatural causes such as old age, starvation, body injury, stress, suffocation, water pollution, diseases, parasites, predation, toxic algae, severe weather, human-induced activities, and other reasons. However, most fish kills result from natural events[9]. The authority attributed the Lake Sebu fish kill event in South Cotabato last January 2021 to the sudden change of weather in the area[10]. Also, extremely high temperature has been identified as a potential agent that could cause fish kill events.

With the identified severe effects of fish kills to the fish farmers, consumers and the fisheries industry, advanced measures and methods must be established to alleviate the adverse effects of fish kill events.

To solve the underlying problem on water quality monitoring, various studies were conducted. The studies presented in [11], [12], [13], [14], and [15] used water sensors and IoT technology to conduct real-time water quality monitoring. However, these studies focused only on the reading and gathering of water parameters.

To solve this, various researchers presented studies that used fuzzy logic to automatically analyze the water quality without the human intervention.

In the study of Caldo and Dedios [16], they categorized the inputs as physical, chemical, and micro biological in the fuzzy inference system to analyze the water quality in Taal Lake. Rana and Rani [17] conducted a MATLAB simulation of a fuzzy logic system to determine the percentage of fish health in freshwater using temperature, dissolved oxygen, and

conductivity. In the study of Ichsan, Kurniawan and Huda [18], they designed a fuzzy logic control based on graphical programming to monitor the water quality in shrimp pond by considering the salinity and turbidity. Hiyunissa, Alam and Salim [19] designed fuzzy logic control system to control microbubble aeration to maintain a desirable value of dissolved oxygen. They only considered the DO and the water temperature. Bokingkito and caparida [20] implemented an IoT-based real - time water quality assessment monitoring system for freshwater aquaculture they considered temperature, pH and turbidity.

Despite the notable results of the studies presented above, their scopes were focused mainly on the assessment of the water used in freshwater aquaculture in general. However, diverse freshwater species have their own tolerance to various water parameters [21] and the desirable ranges of water parameters vary in every freshwater specie.

Hence, this study developed a fuzzy logic-based water quality assessment, specifically focused on Tilapia aquaculture. This study considered pH, DO and temperature as inputs to the fuzzy inference system.

II. The System Model

A. Water Parameters

With the significant profit gained from the aquaculture industry, measures must be taken into consideration to ensure aquaculture's production growth to continuously contribute to the different pattern of supply and demand for fish and fish products. Poor water yields to poor product quality thus this becomes a potential risk to human safety. Also, production is reduced when the water contains contaminants that can impair development, growth, reproduction, or even cause mortality to the cultured species [22].

Massive fish kill events are mostly associated with a sudden change in water composition due to natural causes. Fish kills occur most frequently during the summer when water temperature is high and dissolved oxygen levels are low[23]. Although there is a wide variety of natural and unnatural causes of fish kill incidents, the following are just the identified water parameters that mostly caused sudden massive Tilapia fish kill incidents in the Philippines that need to be monitored in a regular basis.

The first parameter is the water temperature. Water temperature can adversely affect the water condition. Both low and high heat can directly influence other important components that are beneficial in ensuring the health of the marine species [24].

Second is the pH level. The pH concentration of the water can affect the aquatic organisms' health. In freshwaters, inadequate pH levels can hasten the release of metals from rocks and sediments. These eventually affect the metabolism of the fish and its ability to take up water through the gills. Furthermore, low pH can condense the amount of dissolved inorganic phosphorus and carbon dioxide available for phytoplankton during photosynthesis. In contrast, high pH levels can turn the toxic form of ammonia, become more predominant, and the phosphate can rapidly precipitate [22].

And the most important water parameter is the Dissolved Oxygen (DO). Fish needs dissolved oxygen to breathe and perform metabolic activities. Thus, an inadequate level of dissolved oxygen is often associated with fish kill events. Alternatively, precise levels can result in good growth leading to high production yield [22].

B. Desirable Ranges of Water Parameters

The tolerance of diverse aquamarine species to several water parameters varies.

Table I shows the water quality standards for Tilapia Freshwater Aquaculture in the Philippines [21]. It contains the three (3) identified parameters that are crucial in fish kill incidents together with their desirable ranges.

TABLE I. Water Quality Standards for Tilapia Aquaculture in the Philippines

Water Parameters	Desirable Range/s
Temperature	25-32 °C
Dissolved Oxygen (DO)	3-5 mg/L
pH Level	6.5-9

C. Development of the Fuzzy Inference System

There are numerous water parameters that can cause fish kill incidents, however, with the recommendation of an expert[1] in the field of freshwater aquaculture, this paper only focused on the three (3) identified water parameters that caused most of the sudden Tilapia fish kill occurrences. Also, the said expert helped in defining the linguistic values and universe of discourse for each input parameter. Matlab R2018b was used in designing the fuzzy inference system.

In designing the fuzzy inference system of the three (3) inputs, the Membership Function (MF) type used was trapezoidal (1) because there are ranges from a single linguistic value that are equal regarding the degree of membership.

$$f(x; a, b, c, d) = max\left(min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \quad (1)$$

where x is the input value while parameters b and c define the shoulders of the membership function, and a and d define its feet.

In designing the fuzzy inference system of the output, the triangular membership function (2) was used.

$$f(x; a, b, c) = max\left(min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (2)$$

where *x* is the input value while parameters *a* and *c* define the *feet* of the membership function, and *b* defines its *peak*.

The Mamdani Fuzzy Inference System model was used in this paper because of its ability to synthesize a set of linguistic control rules obtained from a human expert and because it has more intuitive and easier to understand rule bases suited to expert system applications [25].

[1]Dr. John Henry R. Centeno III, Fish Health Specialist, Fish Health Laboratory, Bureau of Fisheries and Aquatic Resources, San Mateo, Isabela, Philippines, johnhenrycenteno@gmail.com.

In creating the rules, the logical AND operator was used. The AND operator was used to combine the different antecedents because the value of each input has a significant impact on each other.

And finally, the Center of Area (3) was used for the defuzzification process.

$$CoA = \sum_i \mu(x_i)x_i \; / \; \sum_i \mu(x_i) \quad (3)$$

where $\mu(x_i)$ is the membership value for point x_i in the universe of discourse.

Table II shows the inputs, their linguistic values, Membership Function types and universe of discourse of each water parameter and the structure of the output of this fuzzy inference model.

TABLE II. INPUTS, LINGUISTIC VALUES, MF TYPE AND UNIVERSE OF DISCOURSE

Inputs	**Linguistic Values**	**MF Type**	**Universe of Discourse**
Temperature	Low Normal High	trapmf trapmf trapmf	[12 12 19 28.5] [21.5 25 32 35.5] [28.538 42 42]
Dissolved Oxygen (DO)	Low Normal	trapmf trapmf	[0 0 1.5 3.5] [2.5 4.5 5 5]
pH Level	Acidic Neutral Basic	trapmf trapmf trapmf	[1 1 4 7.25] [5.25 6.5 9 10.25] [7.75 11 14 14]
Output	**Linguistic Values**	**MF Type**	**Universe of Discourse**
Water Quality	Poor Average Normal	trimf trimf trimf	[0 0 0.5] [0.3 0.5 0.7] [0.5 1 1]

Fig. 1 shows the concept of the fuzzy inference model of this study.

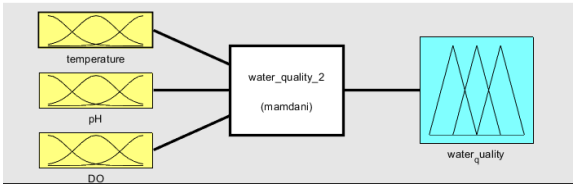


Fig. 1. Water Quality Fuzzy Inference System Model

Fig. 2 to 4 shows the design for each water parameter as shown in their captions.

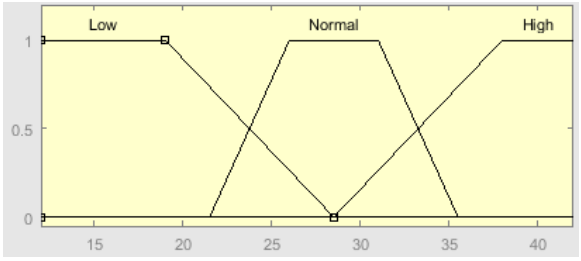


Fig. 2. Temperature Membership Function.

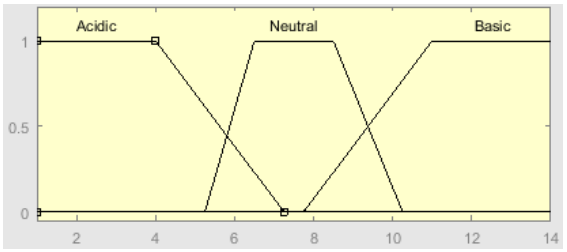


Fig. 3. pH Level Membership Function.

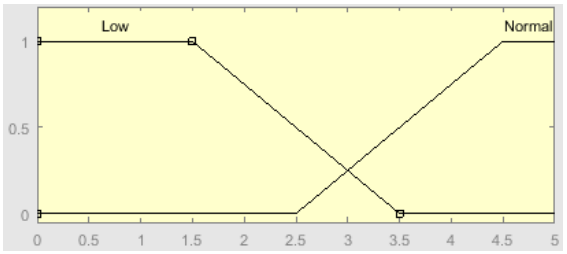


Fig. 4. Dissolved Oxygen (DO) Membership Function.

Fig. 5 illustrates the model for the overall water quality.

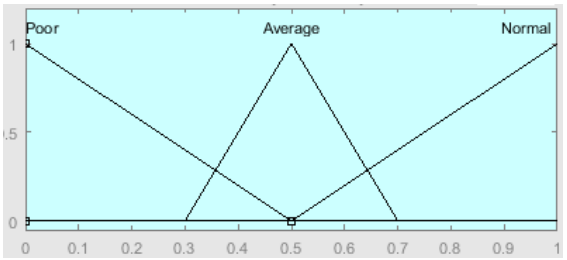


Fig. 5. Water Quality Membership Function.

With the assistance and guidance of the expert in Tilapia aquaculture, eighteen rules were generated for the model.

D. Development of the AquaStat Device

Arduino Uno was mainly used as the microcontroller of the device. The device is equipped with three (3) ATLAS sensors, namely: pH Sensor, dissolved oxygen (DO) sensor, and temperature sensor. Also, this device has a data logger and LCD to display the sensor readings and the overall water quality.

Arduino IDE was used to write the code and the fuzzy logic implementation of this paper.

Fig. 6 shows the actual AquaStat device.

Fig. 7 shows the conceptual model of AquaStat with fuzzy inference system.

Fig. 6. AquaStat Device.

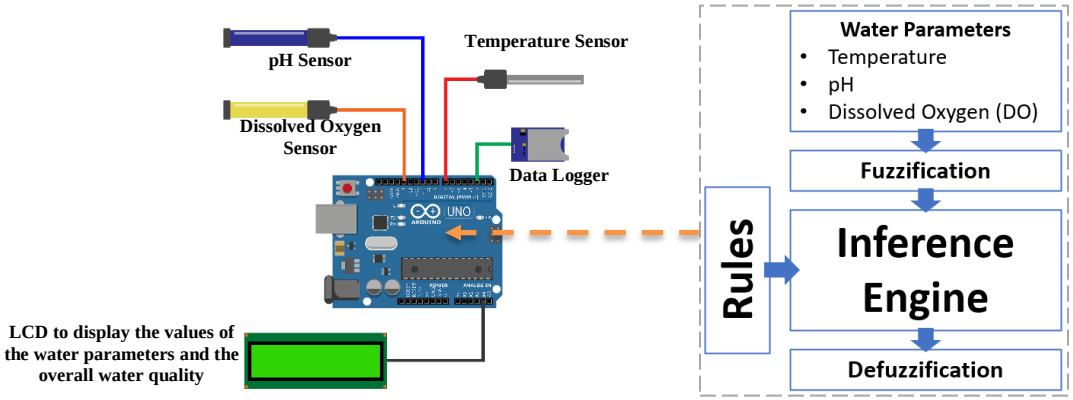


Fig. 7. Conceptual Model of AquaStat.

III. Results and Discussion

A. Accuracy in Gathering Water Parameter Values

The testing of AquaStat was done in five (5) fish cages at the Fish Health Laboratory, Bureau of Fisheries and Aquatic Resources, San Mateo, Isabela, Philippines with the assistance of Dr. Henry Centeno, Fish Health Specialist.

To determine the extent of the accuracy of AquaStat in terms of reading the water parameters, comparison was made against the obtained readings of LaMotte Fresh Water Aquaculture Test Kit which is currently being used in the said laboratory.

Table III shows the obtained pH level of the water in five (5) fish cages.

The obtained pH values from the sample fish cages showed differences having a mean of 0.14. However, since the LaMotte Fresh Water Aquaculture Test Kit is a manual device and only capable of reading pH values with limited and exact decimal values, such differences in decimal values are considered insignificant.

Table IV shows the obtained dissolved oxygen level of the water in the five (5) fish cages.

The obtained DO values from the sample fish cages showed differences having a mean of 0.13. However, since the LaMotte Fresh Water Aquaculture Test Kit is a manual device and only capable of reading DO values with limited and exact decimal values, such differences in decimal values are considered insignificant.

TABLE III. pH Readings using AquaStat and LaMotte Fresh Water Aquaculture Test Kit

Cages	AquaStat	LaMotte Fresh Water Aquaculture Test Kit	Difference
Fish Cage 1	9.14	9	.14
Fish Cage 2	8.21	8	.21
Fish Cage 3	8.17	8	.17
Fish Cage 4	7.67	7.5	.17
Fish Cage 5	7.53	7.5	.03
Mean Difference			**0.14**

TABLE IV. Dissolved Oxygen (DO) Readings using AquaStat and LaMotte Fresh Water Aquaculture Test Kit

Cages	AquaStat	LaMotte Fresh Water Aquaculture Test Kit	Difference
Fish Cage 1	3.43	3.3	.10
Fish Cage 2	2.72	2.6	.12
Fish Cage 3	3.83	3.7	.13
Fish Cage 4	2.61	2.5	.11
Fish Cage 5	2.87	2.7	.17
Mean Difference			**0.13**

Table V shows the obtained water temperature in the five (5) fish cages.

TABLE V. Water Temperature Readings using AquaStat and LaMotte Fresh Water Aquaculture Test Kit

Cages	AquaStat	LaMotte Fresh Water Aquaculture Test Kit	Difference
Fish Cage 1	29.33	29.30	.03
Fish Cage 2	29.27	29.20	.07
Fish Cage 3	28.87	28.80	.07
Fish Cage 4	29.35	29.30	.05
Fish Cage 5	28.91	29.80	.11
Mean Difference			**0.07**

The obtained water temperature values from the sample fish cages showed differences having a mean of 0.07. However, since the LaMotte Fresh Water Aquaculture Test Kit is a manual device and only capable of reading water temperature values with limited and exact decimal values, such differences are considered insignificant.

B. Accuracy in Assessing the Overall Quality of the Water

To further evaluate the accuracy of AquaStat in identifying the overall quality of the water, actual historical data of the water parameters during the regular monitoring activities of the concerned agency were used as inputs for each respective water parameter in the simulation process, Fig. 8.

Fig. 8. Simulation of the Program using Arduino IDE

Table VI shows the results of identifying the overall water quality of the water using the historical data by the Fish Health Specialist and AquaStat. Using the given historical data, the researcher asked Dr. Centeno to evaluate and identify the overall water quality based on his own assessment. The researcher then used the given historical data and inputted them to AquaStat, and the following results were drawn.

Table VI presents the overall analysis of the water quality by the Fish Health Specialist and AquaStat using the actual historical data. The historical data are consisted of values per water parameter. The results of the overall water quality evaluation generated by AquaStat using the collected actual historical data were found to be in consonance with the overall water quality given by the Fish Health Specialist.

TABLE VI. OVERALL ANALYSIS OF THE WATER QUALITY BY THE FISH HEALTH SPECIALIST AND AQUASTAT USING THE ACTUAL HISTORICAL DATA

Water Parameters			**Result of the Assessment by Fish Health Specialist**	**Result of the Assessment using AquaStat**
Water Temperature ((in °C)	***Dissolved Oxygen (in mg/L)***	***pH***		
33.57	2.57	7.7	Poor	Poor
31.43	4.49	9.4	Average	Average
33.20	4.15	7.5	Normal	Normal
33.67	2.78	7.3	Poor	Poor
33.13	2.63	7.4	Poor	Poor
33.21	4.59	7.4	Normal	Normal
32.08	1.12	7.2	Poor	Poor
33.25	1.49	7.5	Poor	Poor
28.45	1.21	7.1	Poor	Poor
30.40	1.53	6.6	Poor	Poor
26.83	4.56	7.8	Normal	Normal
28.19	4.37	8.2	Normal	Normal
25.14	4.45	8.5	Normal	Normal
29.20	3.67	8.0	Normal	Normal
28.77	3.09	7.5	Normal	Normal
28.69	4.97	6.2	Normal	Normal
29.51	4.14	8.0	Normal	Normal
31.30	5.11	8.3	Normal	Normal
32.13	3.66	8.1	Normal	Normal
31.03	4.07	8.9	Normal	Normal

IV. CONCLUSION

Fish kill events in Tilapia aquaculture in the Philippines are undeniably bringing negative impacts in the lives of the Filipino Tilapia fish farmers and the consumers. Undeniably, it is difficult to accurately predict as to when fish kill events will occur, however, having preemptive measures can lessen its devastating effect.

Based on the results, readings of the important water parameters such as pH, dissolved oxygen and temperature using AquaStat showed promising accuracy compared to the device that is currently being used by the subject agency having mean difference of 0.14, 0.13 and 0.07, respectively.

Also, AquaStat can accurately determine the overall quality of the water by using the human reasoning applied to it using Fuzzy Logic obtaining a 100% accuracy.

Hence, the results and the concept of this study can be used in another study on specific freshwater specie.

ACKNOWLEDGMENT

The researcher would like to acknowledge the assistance and help extended by Dr. John Henry R. Centeno III, Fish Health Specialist at the Fish Health Laboratory, Bureau of Fisheries and Aquatic Resources, San Mateo, Isabela, Philippines, johnhenrycenteno@gmail.com.

REFERENCES

[1] Philippine Statistics Authority, "Fisheries Situation Report January to December 2020," 2021.

[2] Bureau of Fisheries and Aquatic Resources, "Philippine Fisheries Profile 2017," 2017.

[3] Philippine Statistics Authority, "Fisheries Statistics of the Philippines," 2017.

[4] Philippine Statistics Authority, "2017 Commodity Fact Sheets," 2018.

[5] Philippine Statistics Authority, "Selected Statistics Agriculture on 2018," 2018.

[6] B. C. Hohls and A. L. Kühn, "Field Guide to Fish Kill Assessments," 2001.

[7] Bureau of Fisheries and Aquatic Resources BFAR-PHILMINAQ, "Managing Aquaculture and Its Impacts: A Guidebook for Local Governments," 2007.

[8] G. S. Jacinto, "Fish Kill in the Philippines — Déjà Vu," Sci. Diliman, vol. 23, no. December, pp. 1–3, 2011.

[9] L. A. Helfrich, E. Specialist, W. Sciences, V. Tech, S. A. Smith, and V. Tech, "Fish Kills : Their Causes and Prevention Fish Diseases and Parasites Collecting and," 2009.

[10] B. Sarmiento, "Lake Sebu fish kill destroys P20M worth of tilapia | MindaNews," MindaNews, Jan. 12, 2021.

[11] M. S. Chavan, V. P. Patil, S. Chavan, S. Sana, and C. Shinde, "Design and Implementation of IOT Based Real Time Monitoring System for Aquaculture using Raspberry Pi," Int. J. Recent Innov. Trends Comput. Commun., vol. 6, no. 3, pp. 159–161, 2018.

[12] B. Shi, V. Sreeram, D. Zhao, S. Duan, and J. Jiang, "A wireless sensor network-based monitoring system for freshwater fishpond aquaculture," Biosyst. Eng., vol. 172, pp. 57–66, 2018.

[13] C. Dupont, P. Cousin, and S. Dupont, "IoT for aquaculture 4.0 smart and easy-to-deploy real-time water monitoring with IoT," 2018 Glob. Internet Things Summit, GIoTS 2018, 2018.

[14] Y. T. Liu et al., "A solar powered long range real-time water quality monitoring system by LoRaWAN," 2018 27th Wirel. Opt. Commun. Conf. WOCC 2018, pp. 1–2, 2018.

[15] B. M. Mathisen, P. Haro, B. Hanssen, S. Björk, and S. Walderhaug, "Decision Support Systems in Fisheries and Aquaculture: A systematic review," no. November, 2016.

[16] R. B. Caldo and E. P. Dadios, "Fuzzy logic control of water quality monitoring and surveillance for aquatic life preservation in Taal Lake," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, 2012.

[17] D. Rana and S. Rani, "Fuzzy logic based control system for fresh water aquaculture: A MATLAB based simulation approach," Serbian J. Electr. Eng., vol. 12, no. 2, pp. 171–182, 2015.

[18] M. H. H. Ichsan, W. Kurniawan, and M. Huda, "Water Quality Monitoring with Fuzzy Logic Control based on Graphical Programming," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 14, no. 4, p. 1446, 2016.

[19] T. Haiyunnisa, H. S. Alam, and T. I. Salim, "Design and implementation of fuzzy logic control system for water quality control," Proc. 2nd Int. Conf. Autom. Cogn. Sci. Opt. Micro Electro-Mechanical Syst. Inf. Technol. ICACOMIT 2017, vol. 2018-Janua, pp. 98–102, 2017.

[20] P. B. Bokingkito and L. T. Caparida, "Using fuzzy logic for real - Time water quality assessment monitoring system," ACM Int. Conf. Proceeding Ser., pp. 21–25, 2018.

[21] BFAR-NFFTC, "Basic Biology of Tilapia," NFFTC Aqua-Leaflet No. 2000-06. 2000.

[22] Philminaq, "Water Quality Criteria and Standards for Freshwater and Marine Aquaculture," Mitigating impact from Aquac. Phillippines, pp. 1–34, 2007.

[23] Aquatic Systems: Lake and Wetland Services, "Florida Fish Kills." 2100 NW 33rd St., Pompano Beach, Florida.

[24] J. K. Buttner, R. W. Soderberg, and D. E. Terlezzi, "An introduction to water chemistry in freshwater aquaculture," NRAC Fact Sheet, no. 170. Northeastern Regional Aquaculture Center, University of Massachusetts, North Dartmouth, Massachusetts, 1993.

[25] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," Int. J. Man. Mach. Stud., vol. 7, no. 1, pp. 1–13, 1975.

Evaluation of Re-identification Risk using Anonymization and Differential Privacy in Healthcare

Ritu Ratra, Preeti Gulia, Nasib Singh Gill
Department of Computer Science and Applications
Maharshi Dayanand University
Rohtak, Haryana, India

***Abstract*—In the present scenario, due to regulations of data privacy, sharing of data with other organization for research or any medical purpose becomes a big hindrance for different healthcare organizations. To preserve the privacy of patients seems like a crucial challenge for Healthcare Centre. Numerous techniques are used to preserve the privacy such as perturbation, anonymization, cryptography, etc. Anonymization is well known practical solution of this problem. A number of anonymization methods have been proposed by researchers. In this paper, an improved approach is proposed which is based on k-anonymity and differential privacy approaches. The purpose of proposed approach is to prevent the dataset from re-identification risk more effectively from linking attacks using generalization and suppression techniques.**

Keywords*—Data privacy; anonymization; differential privacy; re-identification risk analysis; privacy preserving data publishing

I. INTRODUCTION

Due to the advancements in the areas of business intelligence, generally organizations for instance banks, healthcare, health insurance are converted into "data-driven" organizations. These organizations used to apply new mechanisms to analyze a high volume of data. It is the responsibility of the data controller to ensure the user about their privacy and it should be done before publishing the data to a third party. There is no protection of privacy in the original dataset. PPDP (Privacy-Preserving Data Publishing) offered numerous tools and mechanisms to preserve privacy. [1][2][3][4]. Anonymization must be done on the datasets before publishing to various organizations because they may contain personal information. It is well known that personal information can be gathered from these types of records and there are many people who assess the re-identification risk. European Medicines Agency (EMA) recommends an anonymization approach for risk analysis based on qualitative technique and quantitative technique [5].

PPDP process consists of different phases i.e. collection of data; providing storage for collected data; perform anonymization; data publishing after modification and perform data mining process as shown in the conceptual scenario of PPDP described in Fig. 1. There are some persons such as record owner, data holder; data publisher; data recipient, and adversary are involved in this process. The record owner is the entity of record, data holder can be person or organization that holds the data; data publisher is responsible for the publishing of anonymous data; data recipient is any entity that has access to published data and adversary is the entity whose objective is to gather user's information. At the time of the data publishing process, sensitive records may be leaked out. To overcome this problem one possible solution is to modify the dataset. There are many methods for modification of datasets in PPDP [6]. Data anonymizaton is most commonly used to achieve privacy protection in data publishing. Several methods have been proposed to handle the security issues related to datasets. In particular, anonymisation and differential privacy are two techniques that have been used for implementation practically. The k-anonymity used to perturb datasets by generalization and suppression. K-anonymity algorithm is used to preserve user's identity through linking attacks [7]. Differential privacy is also used to prevent privacy by furnishing individuals' personal information ability. However, instead of using k-anonymity's deterministic approach to in distinguishability, differential privacy invokes stochastic in-distinguishability by adding noise or perturbing values. Both k-anonymity and Ɛ-differential privacy suffer from a number of drawbacks. In particular, the curse of dimensionality of adding extra quasi identifiers to the k-anonymity framework results in greater information loss [8]. On the other hand, differential privacy has long been criticized for the large information loss imposed on records. The proposed technique in this paper shows how to overcome these drawbacks by combining k-anonymity and Ɛ-differential privacy, while simultaneously benefitting from their advantages. This paper presents the k-anonymity and differential privacy technique. Both techniques have their own limitations. This can be improved upon in their combination. To implement such a concern is focus of their paper is on re-identification risk analysis.

The rest of the organization of the paper is as follows: Section II provides the literature survey related to anonymization and differential privacy. Section III elaborates the materials and methods used in the paper. Section IV describes the proposed work. Section V presents the experimental details of proposed technique and corresponding results. Section VI concludes the paper.

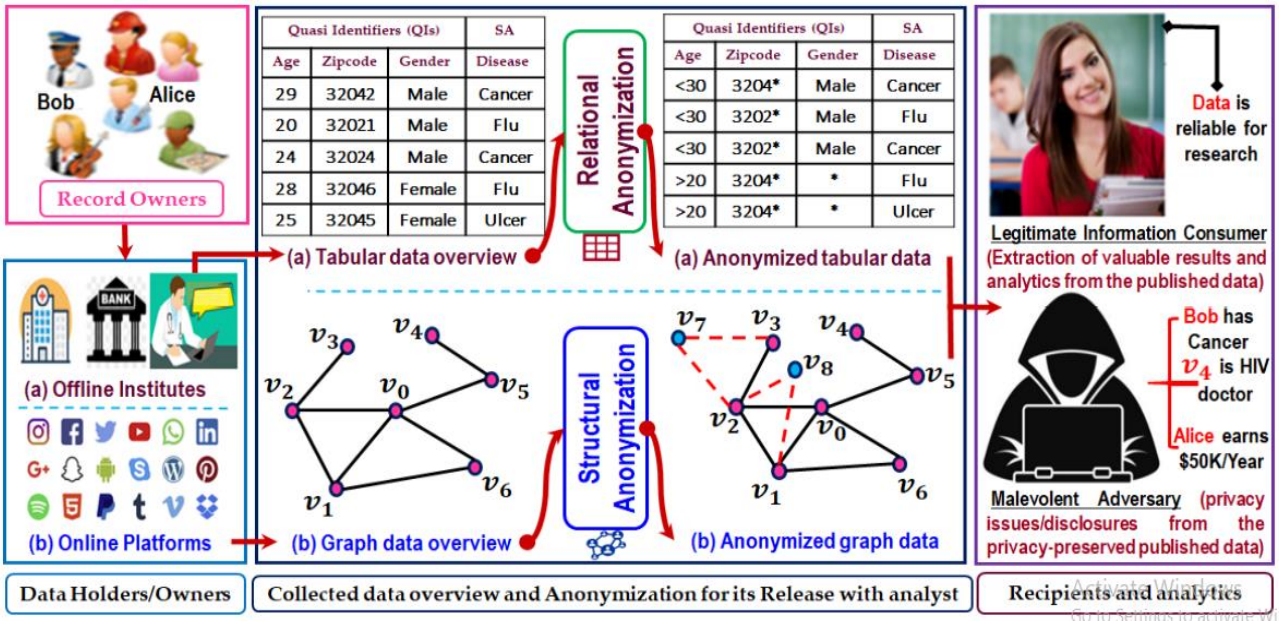


Fig. 1. Privacy Preserving Data Publishing (PPDP) Process [6].

II. Related Work

Protection of sensitive data and extraction of useful information from distributed data is also a challenging task. It is need to preserve the privacy of the before publishing. More than sufficient work has been proposed and implemented in the field of privacy-preserving data publishing. There are several methods used to protect sensitive data. There are various privacy-enhanced mechanisms that are related to the preservation of privacy [9].

Luc Rocher et.al [10] proposed an approach based on the generative copula method. This approach estimated more accurately the probability of anyone to be rightly re-identified.

Boris Lubarsky [11] described a method that proved to be successful even in the heavily incomplete dataset shared. Re-identification can occur due to insufficient anonymization of datasets or combining the datasets. Pseudonym reversal may also be one of the causes of re-identification risk.

Branson, et al [12] have presented a study of testing the re-identification problem. They presented a study through the testing of how a prescribed drug can be subjected to cause of re-identification.

Suman et al [13], introduced a novel technique based on anonymization. The proposed algorithm's performance was measured by using information loss and accuracy. In various experiments, proposed approach provided minimum information loss and maximum accuracy.

Sumana and Hareesh [14] described various anonymization methods in PPDM which are used to provide privacy of the data. Anonymization's main goal is to secure access to personal information and is also used to provide accumulated information.

Vibhor Sharma et.al [15] presented a new Evolutionary privacy-preserving technique in data mining. Whenever data mining is applied to large datasets a number of threats are automatically introduced to privacy. To provide protection to the sensitive data of individuals, data should be masked before it is revealed for data mining.

Marques et.al [16] discussed a complete analysis study on anonymization. A number of techniques of anonymization can be applied to datasets to prevent re-identification risk. They discussed different tools such as ARX, µ-Argus, SDC Micro, and Privacy Analytics Eclipse.

Manoj Kumar Gupta et.al [17] determined various approaches like a generalization, k-anonymity, l-diversity, suppression, shuffling, noise addition, etc. l-diversity is based on the inside group diversity of sensitive attributes. According to the definition of l-diversity, there must be minimum value for each private attribute when each group contains one sharing combination of key attributes. Only then the dataset will be considered as satisfied l-diverse.

P Ram Mohan Rao et.al [18] introduced a novel approach named "Synthesize Quasi Identifiers and apply Differential Privacy" (SQIDP) for privacy-preserving in data mining. This approach was applicable to text data set with 100% data utility.

III. Methods and Techniques Existing

This section highlights the existing techniques and algorithms that are used in proposed technique i.e. Anonymization and differential privacy. These techniques are used to preserve the privacy before publishing.

A. Anonymization

Anonymization is a type of modification technique used to preserve privacy [19]. In data anonymization, sensitive information is either encrypted or removed from the datasets in order to preserve the privacy. There are two methods of anonymization i.e. generalization and suppression [20]. In

Generalization, individual attributes are substituted with an extensive category. Generalization is also a method used for changing categorical attributes and continuous numeric attributes, while suppression means just removing the values of attributes. In this, certain values of the attributes are converted into an asterisk '*'. Various types of attributes are as [21].

Although these types of information may seem very harmless and individually may not present any harm but by linking them from each other, the attackers can misuse can also change the information. In order to hide these original data, there is need to hide and secure these data which may, in turn, present us with another challenge, information loss.

Nowadays, it is common that some of the datasets are openly available for research purpose. To preserve the privacy of shared data, the owner of data can apply different types of anonymization on the datasets. Generalization, suppression, permutation, and perturbation are some examples of anonymization. Furthermore, more than one approach can be applied to the dataset. It proved more beneficial to protect the privacy of data [22]. Therefore, it is necessary to consider the concept of de-identification and re-identification of data. For this purpose, a medical data set has been used that contains the information of some patients. It is depicted in Table I. Here the name attribute is the personal identification attribute; a sensitive attribute is a disease.

TABLE I. DESCRIPTION OF USER'S ATTRIBUTES AND SOLUTION IN ANONYMIZATION

Attribute type	Meaning of Attribute	Solution in Anonymization
Identifying/ Direct	Some attributes like name, mail identity, or aadhar number come under this category. These attributes can certainly recognize the person' personal information	These attributes are removed in anonymization process.
Quasi identify	When one attribute linked with some other attribute caused the disclosure of privacy then those are called quasi identify attributes. For example, age and sex when linked to some other database can easily disclose the person's identity	These attributes are suppressed or generalized in order to preserve the privacy of an individual.
Sensitive	These attributes are crucial and should not be shared. For example. Disease information, salary information should not be shared against any organization.	Mostly do not change for data analyses.
Non-Sensitive	These are the attributes that are publishable publicly because these do not create any problem related to privacy. For example weight, hair color, height, etc.	These are not collected in most cases. If collected, shared as it is.

TABLE II. MICRO TABLE OF HEALTHCARE RECORDS (ORIGINAL)

Name	Zip	Age	Gender	Disease
Wilson	56478	25	M	Heart Disease
Marin	56399	27	F	Blood Cancer
Bob	56789	43	M	Flu Holdon
Emela	56866	34	F	Heart Disease
Peter	56300	24	M	Heart Disease
John	56708	46	M	Prostate Cancer
Boby	56427	33	M	Prostate Cancer

Table II is an example of de-identification. De-identification is the process of altering the dataset to create an alternate use of the dataset so that it is impossible to recognize the identity. De-identification of Table II is shown in Table III, where the field name "Name" is deleted. To provide privacy if the name attribute is removed, then to provide the privacy data can be altered and the altered data is displayed in Table III.

Now the names of patients are not shown in Table III. However, if anyone has access to Aadhar Card Data (as shown in Table IV), it is very easy to discover the information regarding all records. It can be done by joining the two different tables on the common attributes.

TABLE III. HEALTHCARE RECORDS AFTER DELETING THE NAME FIELD (DE-IDENTIFICATION)

Zip code	Age	Gender	Disease
56478	25	M	Heart Disease
56399	27	F	Blood Cancer
56789	43	M	Flu Holdon
56866	34	F	Heart Disease
56300	24	M	Heart Disease
56708	46	M	Prostate Cancer
56427	33	M	Prostate Cancer

These common attributes are known as quasi-identifier. By using the data of Table III and Table IV, an attacker can easily get the information that it is Bob is suffering from a disease of Flu Holdon. So removing the personal information will not be helpful for complete privacy to the data. The method of reversing the de-identification by connecting the identity of the data subject is referred to as Re-identification.

TABLE IV. AADHAR CARD DATASET MICRO DATASET

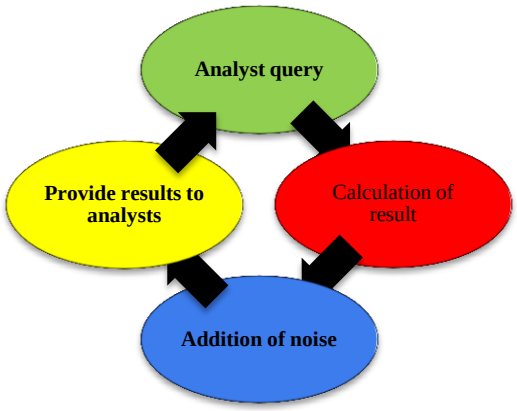
Proposed Name	Zip	Age	Gender
Wilson	56478	25	M
Marin	56399	27	F
Bob	56789	43	M
Emela	56866	34	F
Peter	56300	24	M
John	56708	46	M
Boby	56427	33	M

So in short it can be said that deletion of the personal identification data from relation will not much helpful to protect privacy [23]. To protect privacy first of all personal identification data must be removed and anonymization of the quasi-identifiers is also required.

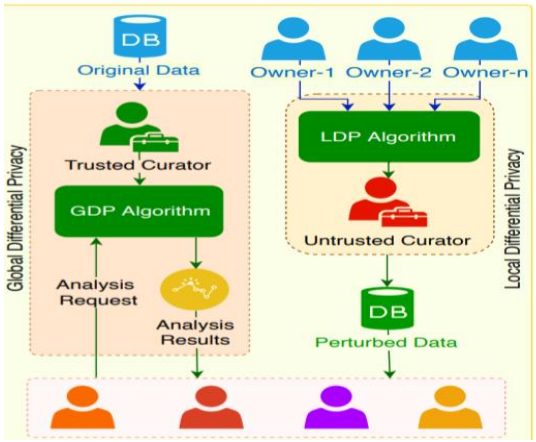
B. Differential Privacy

Differential privacy is also a widely used privacy preservation method. This approach permits the analysts to explore necessary answers from the data repositories that contain sensitive information [24]. In this method, analysts are able to get answers from data stores having sensitive data with secure protection of privacy [25]. In differential privacy, a randomized function R provides Ɛ-differential privacy protection for all data sets named DS1 and DS2. These datasets are differing on at most one data element [26]. This randomized function is such that:

$$Pr [R (DS1) \in S] \leq \exp(Ɛ) \times Pr[\kappa(DS2) \in S]$$



2(a)



2(b)

Fig. 2. (a) Process of Differential Privacy, (b) Global and Local Differential Privacy [16].

Ɛ is the statistical distance, it is use to define the strength of privacy. A lower value of Ɛ means stronger privacy [27]. Different steps of the differential privacy approach are shown in Fig. 2(a). Fig. 2(b) describes the GDP (Global Differential Privacy) and LDP (Local Differential Privacy). A trusted curator recruited in GDP. He can apply gauged noise in order to produce DP (Differential Privacy). The curator should make some practical algorithms or mechanisms that are inappropriate for deep learning. Here the algorithm resides on the server and the original data set has to be uploaded onto the server for training. But in the case of LDP, owners of data modify the data before publishing. There is no need for a trusted curator or any third party to preserve privacy. LDP guaranteed better privacy as compared to GDP. It should be noted that data values are not changed in DP. Here, Users cannot access the database directly. These inaccurate data are sufficient to protect privacy but so small that helpful for the analysts and researchers. Privacy and Utility are not mutually exclusive [28].

IV. Proposed Technique

This paper presents an enhanced privacy –preserving approach based on anonymization and differential techniques. It helps to hide information without abruptly changing the records. The records are k-anonymized as there are k data sets with the same value in each quasi field. To provide anonymization to the original dataset generalization is used. This method is always applied to the quasi attributes [29]. Suppression and generalization techniques are used to provide anonymization. The suppression method is used on quasi attributes in the format of same size intervals. It is done for uniformity in the data set. The proposed enhanced approach tends to solve the privacy issue related to various attacks Generalization is the process through which data can be presented in the form of clustering. The elementary objective of this technique used to collect the links into the cluster and then make a super vertex. Every vertex provides the merged information of the super network. Using this approach, identifying the local data or information is very difficult. To provide protection from re-identification risk, different PPDM (Privacy Preserving Data Mining) techniques [30] are used but the method of anonymity is widely used. This paper proposed the technique of k-anonymity and Ɛ-differential privacy. The proposed method anonymized the data set using a k-anonymity algorithm with k=2 and k=5. The very step first step is to classify the features into sensitive, quasi, and identifiers features. After this, the quasi-identifiers are partitioned into k- quasi on which k-anonymity is applied, and on k- quasi, Ɛ-differential privacy is applied. After this, k-quasi attributes are processed to provide the k-anonymity. After this in the next step differential privacy is applied to the k-quasi attributes. The inspiration to take differential privacy is its stochastic in-distinguishability. Now k-anonymity has applied, an attacker can uniquely recognize the equivalence class. In which any individual's record belongs to that k-quasi. With the help of Ɛ-quasi, it is ensured that the re-identification of records cannot occur.

The proposed method is shown with the help of a flowchart in Fig. 3. It preserves from re-identification risk between equivalence classes. In differential privacy, every equivalent class is considered as a single independent class of an individual's record. In this concept, it is more important to know that differential privacy equivalence class is not the set of attributes. To prevent from re-identification risk records are shuffled.

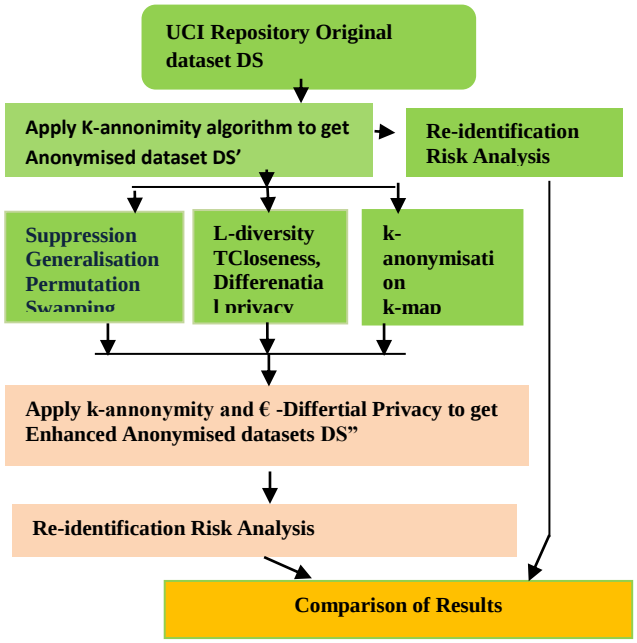


Fig. 3. Flow Chart of Proposed Research Method.

The proposed work is described in Algorithm 1.

Algorithm1: k-ADP (k- Annonymity Differential Privacy)

Input: Original Data set DS

Output: Anonymized data set using k-ADP

Step1. Classify the features (attributes) into quasi, identifiers and sensitive

Step2. Set k-quasi attributes → k-quasi

Step3. Set Ɛ –quasi attribute → Ɛ- quasi

Step4. Apply k-anonymization on k-quasi attributes.

Step5. Apply k-ADP (k- Anonymity Differential Privacy) technique to each equivalence class of k- annonymised dataset

Step6. Now merge k-anonymised records and Ɛ- Differential Privacy records.

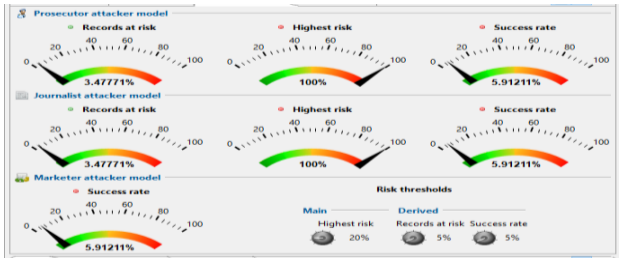
V. Experiments and Results

There are numerous tools and mechanisms for privacy-preservation of datasets. In this paper, anonymization and differential privacy methods are used to provide protection from re-identification risk. From the UCI machine learning repository, Heart dataset is selected for analysis purposes. There are 14 attributes in heart dataset and 2602 records. Out of all attributes, only quasi attributes and sensitive attributes are considered. Here two attributes names as 'age' and 'sex' are considered as quasi attributes and class names as 'result' is considered as a sensitive attribute. Users can directly apply the anonymization method to datasets by using the ARX tool. This tool accepts the files of .csv, .xls, and .xlsv format. Here, k-anonymity with k=2, k=10, and generalization method is selected to perform anonymization on the dataset. Differential privacy is applied to the anonymized dataset. The proposed technique is used to evaate the risk factor of the re-identification. For this purpose, the relationship between k and Ɛ is evaluated. As increases the value of k, the risk is decreased and the risk is decreased with decreasing the value of Ɛ. Now, re-identification risk analysis is done on three datasets i.e. original dataset, anonymised dataset, and enhances anonymised dataset. Experimental results are shown using a tabular and graphical format.

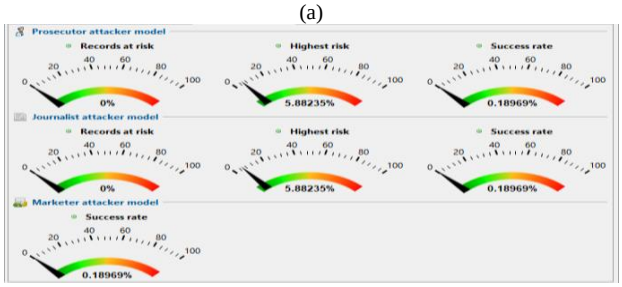
A. Effect on Re-identification Risk

Risk related to privacy can be analyzed using ARX tool [31]. These risks are related to re-identification risk for the prosecutor, journalists and markets attacker. The risk that can be derived from population uniqueness is also included. The impact of data anonymization on the re-identification risk profile for the Heart disease dataset is shown in Fig. 4 and Fig. 5.

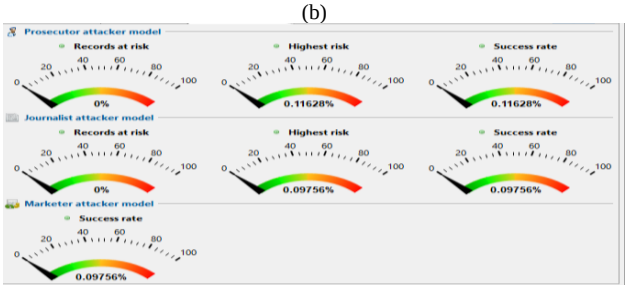
Fig. 4(a) highlights risk of re-identification risk of original dataset at Prosecutor level. Here approximately 3.47% of the total number of records is at risk. The higher risk calculated here is 100%. It means at most all records are at risk in the original dataset. The Success rate is 5.912% in the case of original dataset. At the journalist level, higher risk calculated here is 100%. It means at most all records are at risk in the original dataset. The Success rate is 5.912% in the case of original dataset. It is the same as in the case of the Prosecutor scenario. Fig. 4(b) shows the risk of re-identification of annonymised datasets at the prosecutor level. The highest risk, in this case is 5. 08%. And the effect of the proposed technique is displayed in Fig. 4(c). Here in this case data is purely safe i.e. rate of records at risk is 0% in all scenarios.



(a)



(b)



(c)

Fig. 4. (a) Risk Estimation (Original Dataset), (b) Risk Estimation (Anonymized Dataset), (c) Risk Estimation (Enhanced Anonymized Dataset).

Comparative study of the risk of various attackers of the original dataset, anonymized data set, and enhanced anonymized dataset is given in Table V. Table V lists the risk estimation evaluated at prosecutor level, journalist level, and marketer level. It is depicted that the estimated risk for journalists is higher in the original data set i.e. 33.3% and is lower in enhanced anonymized data set i.e. 0.11%. It can also be noted that estimated Marketer and the Journalist risk are also lowest in enhanced anonymized data set and higher in the original dataset. The detail of various risks is also listed in the Table V. Through the experiments, it is proved that enhanced anonymized data is safer as compared to original data and anonymized data shown in the Fig. 4. The re-identification risk of the original dataset and anonymized dataset is described in Fig. 5 and Fig. 6, respectively.

From Table V, it is stated that the highest Prosecutor risk is higher in the original dataset (100%), and less in enhanced anonymized datasets i.e. 0.11%. Estimated Journalist risk is higher in original dataset (33.30%) and lowers in enhanced anonymized dataset i.e. 0.11%. Estimated marketer risk is higher in original dataset (7.12%), and very less in enhanced anonymised datasets i.e. 0.09%. Re-identification risk estimated in various approaches to the number of records is shown in the following figures.

TABLE V. COMPARISON OF RISK ESTIMATION

Measure	Original dataset	Anonymized dataset	Enhanced Anonymized dataset
Lowest Prosecutor risk	2.12%	0.14%	0.11%
Record at lower risk	4.5%	69.56%	100%
Average Prosecutor risk	7.12%	0.18%	0.11%
Highest Prosecutor risk	100%	.32%	0.11%
Estimated Journalist risk	33.3%	0.32%	0.11%
Estimated Marketer risk	7.12%	0.19%	0.09%.
Estimated Prosecutor risk	33.33%	0.32%	0.11%

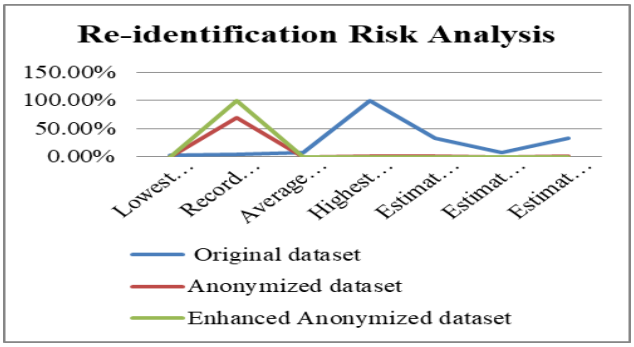
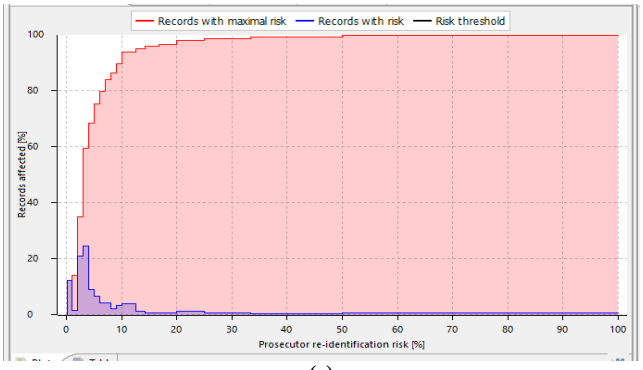
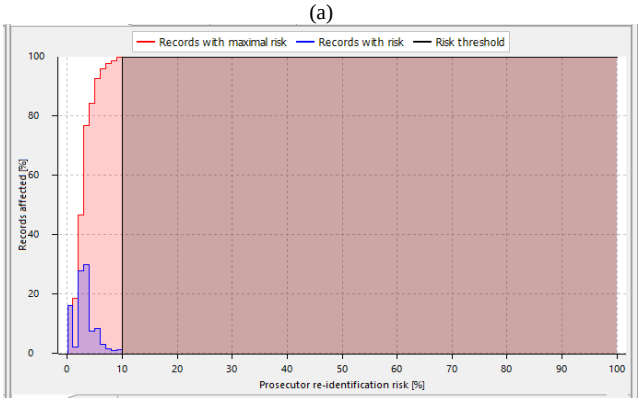


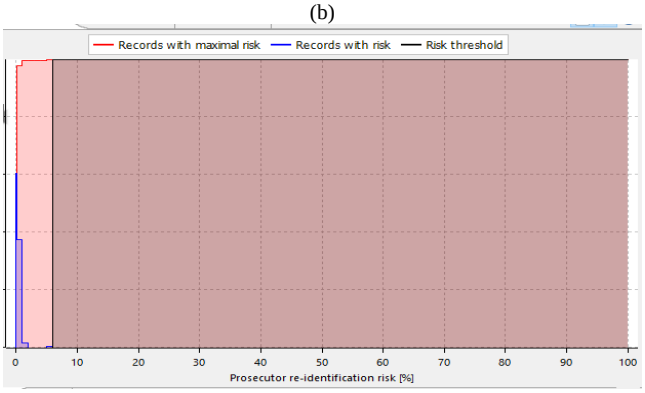
Fig. 5. Re-identification Risk Analysis.



(a)



(b)



(c)

Fig. 6. (a) Re-identification Risk (Original Dataset), (b) Re-identification Risk (Anonymized Dataset), (c) Re-identification Risk (Enhanced Anonymized Dataset).

In the above figures, re-identification risk distribution among the dataset's records is displayed. The calculation of distribution depicted on the input dataset and output dataset. Fig. 6(a) highlights the records with Maximum risk, records of with risk, and risk threshold of the data to prosecutor re-identification risk in percentage. Fig. 6(b) depicted the Maximum risk, Record with risk, and the Risk Threshold of the anonymized dataset at Prosecutor re-identification, and in Fig. 6(c), it is shown that when anonymization with differential privacy is applied on original data set, all three estimations approaches to zero so, the proposed method is much efficient to minimize the re-identification risk.

VI. Conclusion and Future Work

In the era of data sharing, protection of privacy has become an important matter in different organization and in a healthcare industry it is directly concerned with patients. This paper proposed an enhanced anonymized approach to preserve the privacy of patients' data. To preserve the privacy, a proposed technique has been implemented on the dataset related to the heart disease. In this paper, anonymization (K-anonymity) and differential privacy approaches are used to provide privacy to the dataset. Through various experimental results, it is proved that an anonymized dataset achieved more security. The re-identification risk in a modified dataset is very much less as compared to the original dataset. In future, different classification algorithms would be applied to the anonymized dataset to measure the accuracy, execution time, kappa-static, etc.

Acknowledgment

The authors are grateful to the UCI Repository for providing the dataset and also thankful to all members of the Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India for their kind support.

References

[1] Deepak Narula, Pardeep Kumar, Shuchita Upadhyaya, "Evaluation of proposed amalgamated anonymization approach", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 16, No. 3, pp 1439-1446, ISSN: 2502-4752, December (2019). http://doi.org/10.11591/ijeecs.v16.i3.

[2] Gregory E. Simon, et.al, "Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records", The Journal for electronic Health Data and Methods, EGEMS (Wash DC). 29;7(1):6, 2019. Doi 10.5334/egems.27.

[3] P Raje ndra Prasada, Tryambak Hirwarkarb, "Efficient Model for Privacy Preserving Classification Of Data Streams". Turkish Journal of Computer and Mathematics Education Vol.12 No.2, pp. 1475 -1481, (2021).

[4] Ritu Ratra, Preeti Gulia, "Privacy Preserving Data Mining: Techniques and Algorithms", International Journal of Engineering Trends and Technology, Volume 68 Issue 11, ISSN: 2231 – 5381, pp. 56-62, (2020). DOI:10.14445/22315381/IJETT-V68I11P207.

[5] Anastasiia Pika, et.al, "Privacy-Preserving Process Mining in Healthcare", International Journal of Environmental Research and Public Health", ISSN: 1660-4601, pp 1-28, (2020); https//:doi:10.3390/ijerph17051612.

[6] Abdul Majeed, Sungchang le et.al., "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey" , IEEE Access, volume 9, pp 8512- 8545, (2021). https://doi.org/10.1109/ACCESS.2020.3045700.

[7] Can Eyupoglu, Muhammed Ali Aydin, Abdul Halim Zaim and AhmetSertbas "An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques". www.mdpi.com/journal/entropy, pp 1-18, (2018).

[8] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman, "Local Differential Privacy for Deep Learning", IEEE Internet of Things Journal, Vol. xx, no. xx, arXiv:1908.02997v3[cs.LG] 9 Nov 2019. https://doi.org/10.1109/JIOT.2019.2952146.

[9] Kunwar Singh kushwah and Abhay Panwar, "A Privacy Preservation Technique Using Machine Learning Technique", International Journal of Engineering and Innovative Technology (IJEIT). pp 3445-3454, (2015).

[10] Luc Rocher, Julien M. Hendrickx and Yves-Alexandre de Montjoye., "Estimating the success of re-identifications in incomplete datasets using generative models". Nature Communications, pp 1-9, (2019). https://doi.org/10.1038/s41467-019-10933-3.

[11] Boris Lubarsky, "Re-identification of "Anonymized Dat Georgetown Law Technology Review, Vol 1:1, pp 202-213. (2018).

[12] Branson et al. , "Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations", pp.1-9, (2020). https://doi.org/10.1186/s13063-020-4120-y.

[13] Suman Madan and Puneet Goswami, "Adaptive Privacy Preservation Approach for Big Data Publishing in Cloud using k-anonymization", Recent Advances in Computer Science and Communications. Volume 14, Issue 8, pp 2680-2690, (2021). https://doi.org/10.2174/2666255813999200630114256.

[14] Surname, et al., "Information theoretic-based privacy risk evaluation for data anonymization", Journal of Surveillance, Security and Safety, 2021:2:83-102, (2021). https://doi,org/10.20517/jsss.2020.20.

[15] Vibhor Sharma, Dheresh Soni, Deepak Srivastava and Dr. Pramod Kumar.., "A Novel Hybrid Approach of Suppression and Randomization for Privacy Preserving Data Mining", EEO. 2021; 20(5): pp 2451-2457. (2021). doi:10.17051/ilkonline.2021.05.267. https://doi.org/10.1177/2378023121994014.

[16] Marques, J. and Bernardino, J., "Analysis of Data Anonymization Techniques.", In Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD, pages 235-241 ISBN: 978-989-758-474-9, (2020). https://doi.org/10.5220/0010142302350241.

[17] Manoj Kumar Gupta and Abhishek Gupta, "A hybrid-security model for privacy-enhanced distributed data mining", Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx, pp 1-13, (2020).

[18] P Ram Mohan Rao, S Murali Krishna and A P Siva Kumar, "Novel algorithm for efficient privacy preservation in data analytics". Indian Journal of Science and Technology, ISSN Print: 0974-6846 Electronic: 0974-5645. pp. 519-526, (2021). https://doi.org/10.17485/IJST/v14i6.1773.

[19] S Kumaraswamy , Manjula S H , K R Venugopal, "Secure Cloud based Privacy Preserving Data Mining Platform", Indonesian Journal of Electrical Engineering and Computer Science Vol. 7, No. 3, pp830,-838, September 2017. https://doi.org/10.11591/ijeecs.v7.i3.

[20] Alpa Shah and Ravi Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey", International Journal of Computer Applications (0975 – 8887) Volume 137,No.12, pp 40-46, (2016).

[21] D. Kavitha, "A Survey on Privacy Preserving Data Mining Techniques". International Journal of Computer & Mathematical Sciences IJCMS ISSN 2347 – 8527 Vol. 7, Issue 2. pp. 160-169, (2018).

[22] Desmond Ko Khang Siang, et.al, "Comparative Study on Perturbation Techniques in Privacy Preserving Data Mining". International Journal of Innovative Computing 8(1), pp27-32, ISSN 2180-4370, (2019).

[23] Nurislam Tursynbek, Aleksandr Petiushko, Ivan Oseledets , "Robustness Threats of Differential Privacy", arXiv preprint arXiv:2012.07828, 2020 - arxiv.org, Aug (2021).

[24] Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu "A Comprehensive Survey on Local Differential Privacy", Security and Communication Networks, Article ID 8829523, 29 pages, (2020). https://doi.org/10.1155/2020/8829523.

[25] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia, "The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)", Communications of the ACM, Vol. 64 No. 7, Pages 33-35, July 2021, https://doi.org/10.1145/3433638.

[26] Mathew E. Hauer1 and Alexis R. Santos-Lozada, "Differential Privacy in the 2020 Census Will Distort COVID-19 Rates", Socius: Sociological Research for a Dynamic World, Volume 7, pp. 1–6, (2021).

[27] Revathy Swaminathan and T. Arun Kumar, "Survey paper on privacy preserving data mining", International Journal of advanced Research, pp 1120-1127, ISSN: 2320-540, (2016).

[28] Jyothi Mandala, Pragada Akhila, Vulapula Sridhar Reddy, "Integrated Reinforcement DQNN Algorithm to Detect Crime Anomaly Objects in

Smart Cities", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, (2021).

[29] Fabian Prasser and Florian Kohlmayer, "Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool", Gkoulalas-Divanis, Aris, Loukides, Grigorios (Eds.): Medical Data Privacy Handbook, Springer, November. ISBN: 978-3- 319-23632-2. (2015).

[30] Preeti Gulia, Hemlata, "Privacy preserving data mining of vertically partitioned data in distributed environment-an experimental analysis", Journal of Theoretical and Applied Information Technology, Vol.96. No 10, ISSN: 1992-8645, pp 2973- 2987, (2018).

[31] Jaap. Wieringa, et.al, ''Data analytics in a privacy-concerned world,'' Journal of Business Research., Volume 122, pp. 915–925, (2021).

Implementation of QT Interval Measurement to Remove Errors in ECG

S. Chitra
Research Scholar
School of Computing Sciences
VISTAS, Chennai, India

Dr. V. Jayalakshmi
Professor
Department of Computer Applications
VISTAS, Chennai, India

***Abstract*—Wireless Body Sensor Network (WBSNs) are devices that can be ported with different detection, storage, computer, but also communication capabilities. Interfacing was beneficial whenever information was collected by many sources, which may lead to erroneous sensory information. During this paper, an information nuclear fission Ensembles technique for working raw healthcare information through WBSNs during ambient cloud computer settings as described. Monitoring data were collected through various instruments and combined to provide statistics on high movements. The simulation was conducted using the low-cost Internet of Things (IoT) surveillance system on chronic kidney disease (CKD). Biosensors have been used in healthcare surveillance systems to record health problems. Patients with CKD would benefit from the developed surveillance system, which will facilitate the early diagnosis of the predominant diseases. This merged information was then sent into using the Aggregation algorithm can forecast premature cardiac illness and CKD. These groups were housed within a Cloud processing context; therefore these forecasting calculations were distributed. Another lengthy practical investigation backs that system provides application, while those findings were encouraging, with 98 percent efficiency whenever the height of that tree was equivalent with 15, total amount if estimation methods are 40, while the overall predicting job was based upon 8 attributes. We compute a mean square ECG waveform from all available leads and use a new technique to measure the QT interval. We tested this algorithm using standard and unique ECG databases. Our real-time QT interval measurement algorithm was found to be stable, accurate, and capable of tracking changing QT values.**

***Keywords*—e-Health; QT interval; GPU; ECG signal; CKD; IoT**

I. Introduction

In various ways, clinical wearing gadgets differ from overall technologies. Consumer engagement is often far very restricted. The volume of its CPU, information memory, & energy endurance is typically used to make an efficient market [1]. Signals analysis was emphasized for this product at the level which was rarely seen in overall worn computer technologies. Ultimately, on aspects such as confidentiality, dependability, governmental laws, & company's lawful obligation, wearing clinical systems must meet higher standards [2]. Researchers shall concentrate upon clinical wearing gadgets throughout the study. Wireless Body Sensor networks (WBSNs) were changing the face of medical technologies [3]. Worn gadgets having devices that could identify physiologic indicators, portals that facilitate internet connection, & back- and company front computers in storing, analyzing, & presenting data were all part of such systems [4-6]. WBSN& smart gadget technologies aren't quite innovative.

Such transportable monitoring gadgets can help throughout their treatment monitoring chronic illnesses like cardiovascular assaults, bronchitis, and hypertension, as much as enabling monitoring collecting vital facts about such patient's individual wearer's anatomy. Electrocardiogram, blood oxygenation saturation, breathing, & physical obesity may all be monitored pulses [7]. The clinical assist gadget was intended to give lengthy help to disabled people. Regarding individuals undergoing rehabilitative, the portable recovery support gadget integrates surveillance & healthcare support capabilities [8]. This monitoring function improved that child's recuperation by allowing them to escape potentially dangerous situations including risks. Furthermore, throughout recovery, this healthcare assistance gadget may assist in transitional limitations.

Its portable product's detectors monitor physiologic signals of its physique [9]. Non-invasive, dependable, small, portable, changeable, & ready to integrate into a gadget are all desirable qualities. Its monitor's selection is dependent upon its intended usage [10]. When its intended purpose was the cardiac tracking device, its device's characteristics, like sensitivity and sample frequency, must be determined [11]. Next researchers must pick any of those detectors which best fit our needs.

The highlights of this work are as follows:

1) Data from various sensors is combined to provide higher-quality data that is fed into ensembles for heart disease and CKD prediction.

2) The ensembles are placed in a fog computing environment, and the separate forecasters' findings were integrated into providing an overall cohesive outcome.

3) Within such cloud computer context, researchers explore using innovative kernels randomized woodland towards its forecasting problem. Exponential randomized forests have been found to significantly outperform commonly used compositions throughout terms of forecasting ability.

II. Literature Survey

The portable gadget, by speaking, offers restricted energy & computing performance. Because of this constraint, the choice must be made on both types & quantities of information that should be collected, saved, & communicated. A portable healthcare equipment's primary goals were able to assist physicians or caretakers via delivering relevant data about the client's condition [12]. This would neither suggest that ascending sensory material must be ignored; rather, basic information must be modified to give better valuable facts in judgment [13]. As this result, researchers must examine which data should be sent to a physician or caretaker, as well as whether to analyze this to receive valuable data. Among the most important technology for wearing healthcare equipment was communications. Connectivity among sensing boards & back-end systems, as well as the linkage among a back-end scheme & its medical practitioner, are often involved [14]. There exists a standardization problem for considering when it comes to device-to-device compatibility.

The primary purpose of such a sensing layer was acquiring physiologic data on this target [15]. Wireless and cable connections would be being used to send the information beyond an information layer, where they would either be saved or exchanged. On this applications layer, these impulses were examined & recreated as treated information. Instruments in any surveillance system were often connected to the destination, whether could be either a physician or the gadget operator [16]. Sensor's elements, a device with cameras pretreatment like is filtration & compression, & communications among both sensors & information plane are all included within a monitors sector [17]. The instrument parts were generally chosen based on the anticipated application. This sensor element must detect its glycemic levels when its doctor wishes to track a concentration of sugar within the patients.

Because of many sensors the shortage of resources substantial computing skills, very basic pretreatment methods like filtration & compression were commonly included. Processing could make a gadget better personality & functional particularly while its networking connectivity was interrupted, however, it must be done with at least quantity of information deformation possible [18]. To communicate basic information for an information plane sensing layer, each sensing plane typically includes a minimum single connected and wirelessly communications link. Challenges upon that sensing plane include a sensing vehicle's lifespan & physical architecture, customer acceptance, confidentiality, & safety issues [19]. This sensing plane's key characteristics were immediately tied with a sensing gadget. Every sensing gadget must assess when long it will continue to gather its data. Whenever a detector is to detect movement, this must be triggered whenever that movement happens. When a sensor is required to capture ECG data, it must be operational over the length when duration.

This information plane serves for a system bridging even though that was placed among its sensing planes & both applications layers. Although that sensing plane and applications planes deal with precise focused information, its information planes must avoid being too generic [20]. This must contain all appropriate connections that facilitate gathering information via many sensing units. Any program might be required to merge and recycle information as the result of a constraint. Because of its sensitivity to physical information, it is necessary must offer such connections proper protection of information accessibility [21]. Because an information plane has 2 routes of transferring information, one is sensing planes and another for the applications planes, all regulations on obtaining information across every platform must be distinct.

Generic networking such as the internet or mobile networking such as GSM, UMTS, CDMA, & WiMAX is used to communicate between an information plane towards the applications planes. An applications layer, by broad, arranges their unique approach of presenting, viewing, & analyzing information across a generalized substrate like computers, tablets, or cellphones [22]. As a result, information interchange becomes a principal source of contact among information planes & an applications layer. Just at software planes, these were two types of information sharing methodologies: sealed information swapping & accessible information swapping. Holding apps often employ connections internet technology can transport data into another informational level.

III. Proposed Method

The echocardiogram was used to examine such electrical motion within the heart (ECG). This detects electronic signals produced from heart muscle polarization & relaxation & converts them to a waveform. This is another critical tool for cardiovascular research & diagnostics. A P waveform, a QRS complicated, a T tidal, and a U generate buzzing swarming up a normal ECG trace of a heart cycle. The P & T phases typically exceed this U waveform, which was typically undetectable. This Interphase was my attention since it has proven proved to be an excellent predictor of existence cardiovascular problems. A timing differential from commencement to that QRS or a conclusion with a T wave was used to calculate it. Signals separation was required to estimate the QT duration, however, that was difficult owing to the peculiarities of ECG constituent waveforms. A QT duration must be adjusted when examining QT readings across the period at varying cardiac rates because it fluctuates on cardiac speed. In generally, Bazzett's equation (1) was commonly employed to calculate an adjusted QT duration, abbreviated QTC:

$$QT^c = \frac{\|QT\|}{\sqrt{\|RR\|}} \quad (1)$$

wherein |RR| denotes the duration of the R-R intervals across its present R maximum & its preceding R high, & |QT| denotes the distance at that start of Q & that conclusion of T. To track the QTC upon a beat-by-beat basis, researchers must separate all calibration events for each pulse, including for its QRS start, T discrepancy, & R maximum. This proposed approach demonstrates an efficient approach that divides every calibration spot & computes QTC enabling heart rate assessment shown in Fig. 1.

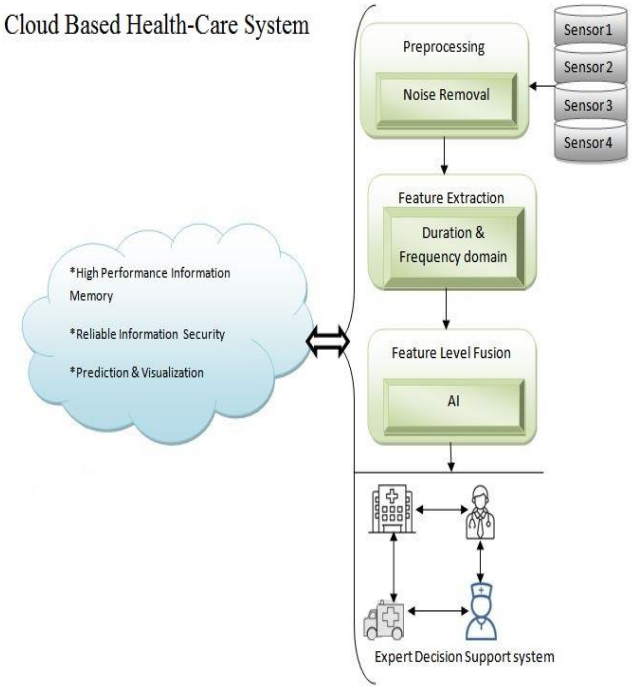


Fig. 1. Flow Diagram Cloud-Based Health Care System.

Researchers require a previous understanding of underlying structures in relevance for locating calibration sites given any series of data. Such was referred to as prototypes and they were standard structures that researchers were looking for. Creating the framework using numerous streams is proposed within this technique. Si denotes the ith individual sample from the i[th] channel (j).T is a pattern that may be described by

$$T_s = |T_1, T_2, \dots, T_n| \tag{2}$$

$$T_x = |S_x(y-h), S_x(y-h+1), \dots\dots, S_x(y), \dots\dots, S_x(y+h)|^T \tag{3}$$

when n denotes the number of networks employed throughout this study & 2h + 1 denotes the overall duration of a proposed system. To reduce the influence of disturbance, researchers use the Gaussian functional to create company center balanced masks. Where U_x be the 1D Gauss weighting vectors of length 2k + 1 for such ith channels.

$$(n) = exp\left(-\frac{1}{2}\left(\frac{|i^2-h^2|}{\partial^2}\right)\right) \tag{4}$$

$$U_x = |h_1, h_2, .. h_2|^T \tag{5}$$

The Gaussian value U can be defined as

$$U = |U_1, U_2, \dots, U_n| \tag{6}$$

Where U_1=U_2=⋯=G_n. The final template T author use in the analysis can be defined as

$$T = Mask(U, T_s) \tag{7}$$

where, Mask (E, F) returns the elements-wise product of E & F.

IV. Result and Discussion

Several sources like disturbance could readily compromise ECG readings. This preparation stage's goal was to filter out interference which is not as legitimate ECG data. One of those most prevalent ECG data the foundations started sliding due to irregularities. Organization drifting & higher frequencies noisy artifacts could be reduced using band-pass filtration. Because that bulk of ECG impulses were found as in inferior range below 100 Hz, our proposed approach uses the band-pass filters having cut-off wavelengths between 0.5 Hz & 100 Hz. Another pseudo-periodic output includes an ECG pulse. Each pulse must be segmented to a beat-by-beat assessment. This QRS Sequence was the most prominent output of every ECG rhythm. It might serve as a useful divider between neighboring rhythms. Each calibration position must be determined among both earlier & later R spikes, & its R-R separation must be calculated using Bazzett's algorithm to determine an adjusted QT period. Because of that, every QT research requires locating the QRS complicated.

Researchers currently know the location of every calibration site, its commencement of such QRS, & its dispersion of T waves. Such structures were depicted within these designs, which might clarify those fiducially locations in the middle of all those frameworks. Our proposed approach identifies its most matched site regarding the provided pattern during a detecting stage. This was just a single matched spot in such a beating data that was a signal across successive R spikes, according to a property in this ECG sensor. During heart rate assessment, our proposed approach uses the outcome of a high point sensing phase as just a beating divider. Where R = {r1, r2,...,rn} represent a maximum identification result. When addressing any boundaries within the proposed approach, the enlarged signals would be being used up towards the duration of such a specified template.

$$E_x = \left|S_x\left(r_h - \frac{z}{2}\right), S_x\left(r_h - \frac{z}{2} + 1\right), \dots, S_x\left(r_{h+1} + \frac{z}{2}\right)\right|^T \tag{8}$$

$$E = |E_1, E_2, \dots, E_n| \tag{9}$$

Whereas n specifies this same number many channels will get studied and w indicates its pattern length.

To get that greatest fitting place when matched to the particular pattern, researchers must evaluate the overall closeness of every place towards the supplied framework. The approach uses a Pearson correlations value, which is described as the measure of the closeness among such signals & the particular framework shown in Fig. 2.

$$\boldsymbol{r_{ij}} = \frac{\sum_{x=1}^{m}(I_x-\bar{I})(J_x-\bar{J})}{\sqrt{\sum_{x=1}^{m}(I_x^2-\bar{I}^2)}\sqrt{\sum_{x=1}^{m}(J_x^2-\bar{J}^2)}} \tag{10}$$

That slicing signals Bs for such a specific position k, that is the identical lengths as any templates, may be described by

$$E_{x,s} = |E_x(h), E_x(h+1), \dots, E_x(h+z-1)|^T \tag{11}$$

$$E_s = \left|E_{1,s}, E_{2,s}, \dots, E_{n,s}\right| \tag{12}$$

when, w is that width for a specified pattern & n denotes the number of streams under consideration. Every one of those indicators would be explored by computing Pearson terms of

interaction using company particular templates following component sequential multiplying using the Gaussian masking, as specified by Eq. (11).

A. Decision Rule

Researchers must explore those locations which match all requirements connected towards the features to validate the overall authenticity of a sought fiducial location. Researchers currently get knowledge upon that broad-spectrum between periods based upon its limits on the individual signals. These next checks are used in this technique to demonstrate the integrity of a supplied collection in Equation (13).

$$QT_c^x = \left|r_x, r_{x+1}, Q_{on}(x-1), T_{off}(x)\right| \quad (13)$$

Researchers could derive ith adjusted QT intervals from one pair of QTic.

Several people have a heartbeat with fewer than 220 beats per minute. The usual limit of an adult's heartbeats is 60 - 100 beats per minute. Youth individuals, on the other hand, had quicker heartbeats than elders. A spectrum in adult heartbeats is depicted on this page. Our proposed method limits your pulse rhythm spectrum between 40-200 beats per minute. The offset must be found before Qset for a one beats output. This proposed approach excludes all items submitted in evaluation if a condition was never satisfied with this provided data. Several studies had attempted to demonstrate a rectified QT interval's potential length. This spectrum must remain around 0.3 & 0.6, according to experiments. For being proven like a legitimate database, any submitted information source should likewise fulfill these conditions.

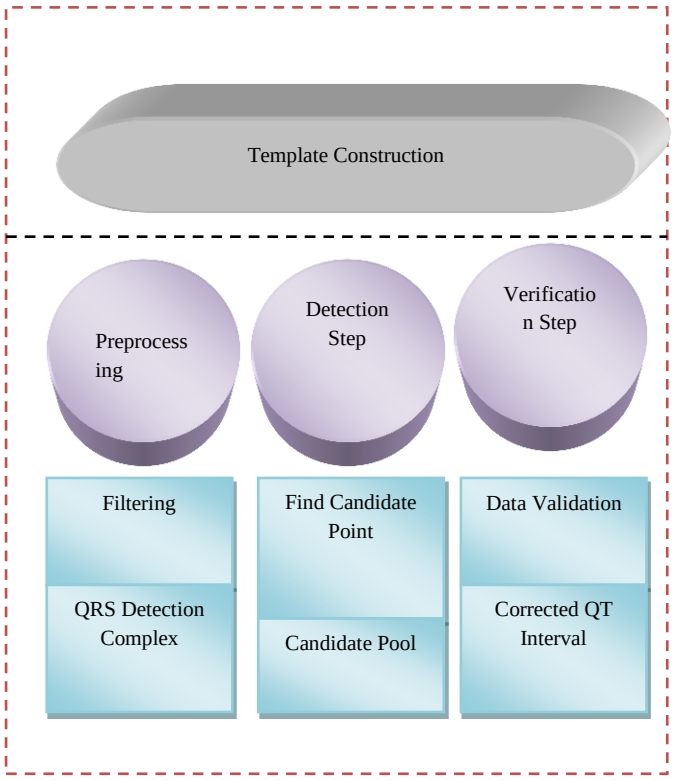


Fig. 2. The Proposed Subsystem Approach.

B. Testing

Researchers provide my method for implementing the recommended method. This architecture for the technology is depicted in Fig. 3. This technology was divided into two sections: service and user. Because this was an internet program, its user sides were built up using web interactions that communicate through HTTP.

A Global Standards of Healthcare Photos & Associated Data was Computerized Imagery & Telecommunications for Health. This specifies codecs of healthcare pictures X-rays, CT examinations, mammograms, and geomagnetic receptor tomography are among examples (MRI) that may be shared while maintaining all information & clarity required for therapeutic application. Frequency information, inspired are electrocardiographic & circulatory information, could be altered within the DICOM standard, as per the DICOM supplemental. The DICOM translator was the data center application. Researchers implement business DICOM standards on my program to allow customers can manipulate electrocardiographic data for sources towards a computer while maintaining compliance using records of commonly employed recording equipment.

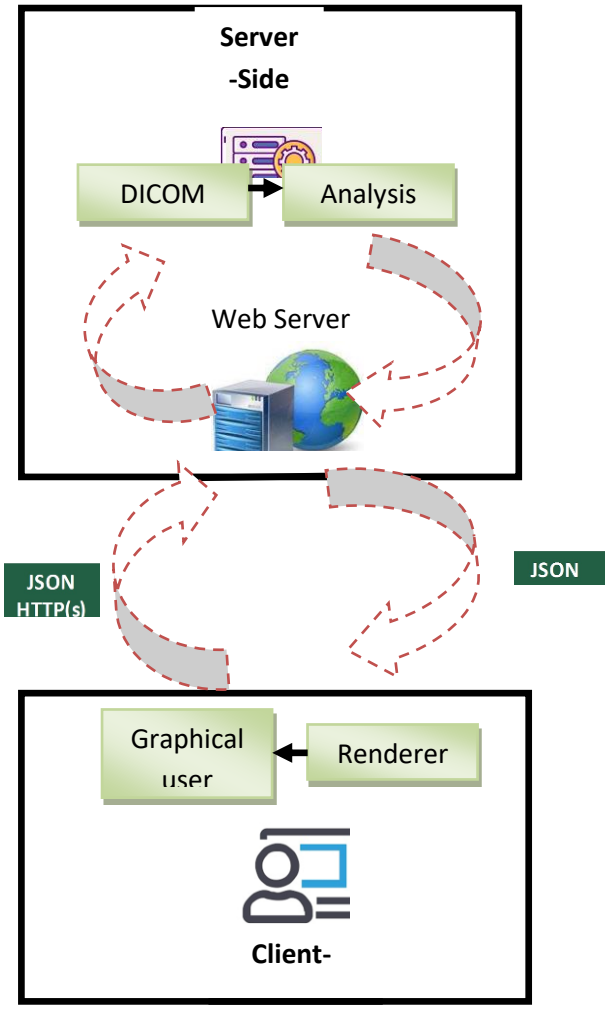


Fig. 3. Design of the Webserver of QT Research.

As per Fig. 2, this information was prepared as DICOM on the user end and later communicated towards the service end via GUI. Because JSON via HTTP was straightforward & quick to deploy, researchers utilize this as an information exchange mechanism. All transferred information was converted into amounts only at the DICOM decoder via interpreting information supplied within the DICOM standard. That digitized output may be used as an output for the proposed method in 5.2, which was performed within the research engines' service element. Some findings of an operation were removed towards such internet host, which converts these into JSON-formatted information & sends this to a user. Researchers used python 2.7 & the Tornadoes internet browser here on the service end, &angularjs& HTML5 just on the consumer edge.

Upon that consumer end, we'll have to create multiple graphical customer interface webpages to interact with a customer & view inputs & process information.

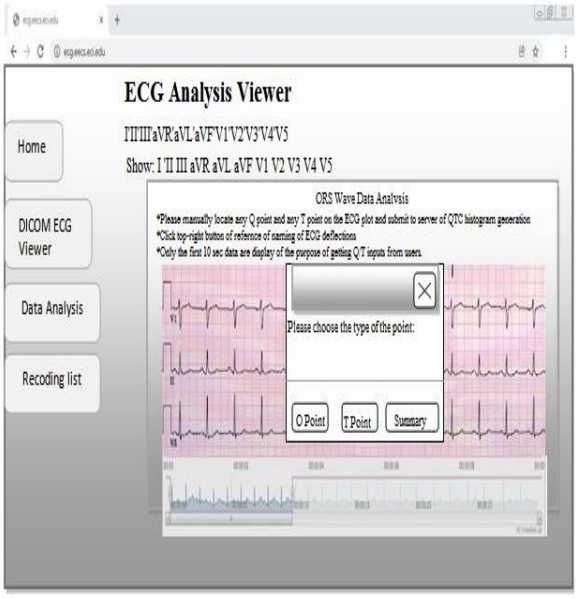


Fig. 4. Selecting of Fiducially Elements in Creating Patterns there in GUI.

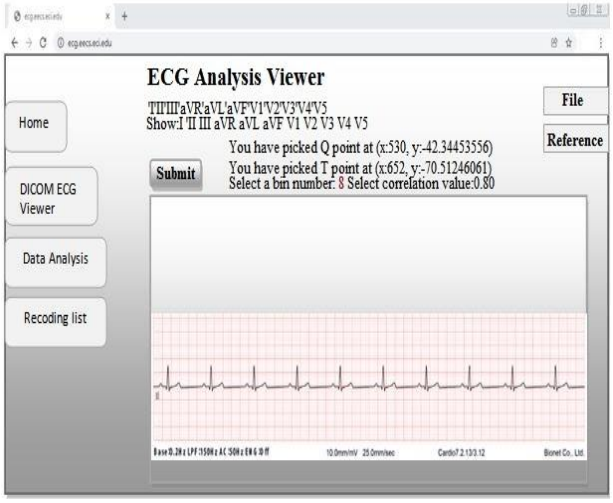


Fig. 5. Initializing the Attributes.

This GUI interface for selecting fiducial locations is shown in Fig. 4. These equations of QRS start & T Distortion of the surges were dynamically developed thanks to customer input here on location. Researchers must establish starting settings as shown in Fig. 5 following constructing algorithms of looking to every fiducial location. These R spikes are found using individual streams using a proposed technique. As a result, consumers can select so that channels are covered for peak monitoring. An additional variable, thresholds, was a minimal number that must be fulfilled for modeling & indicators can remain comparable. When the overall crossing metric rating for relationships of the potential fiducial spots is lower over a criterion, the proposed method excludes them within the evaluation. A quantity of more bits of such QT measurement results was likewise determined. QT assessment outcome would be shown on your GUI as a histogram-like dispersion.

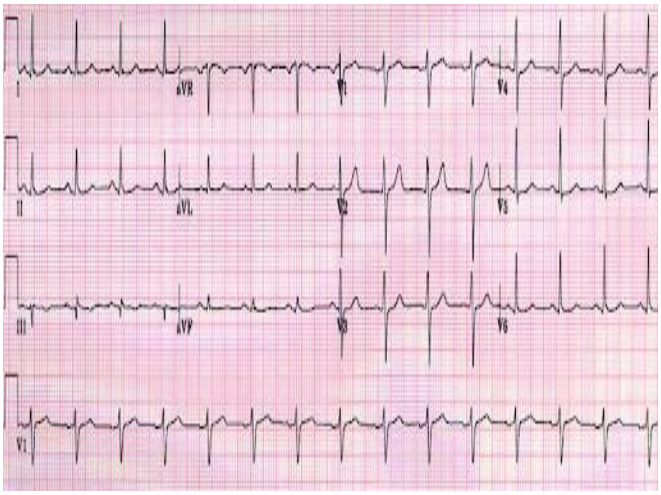


Fig. 6. Image of an ECG in the Typical ECG Style.

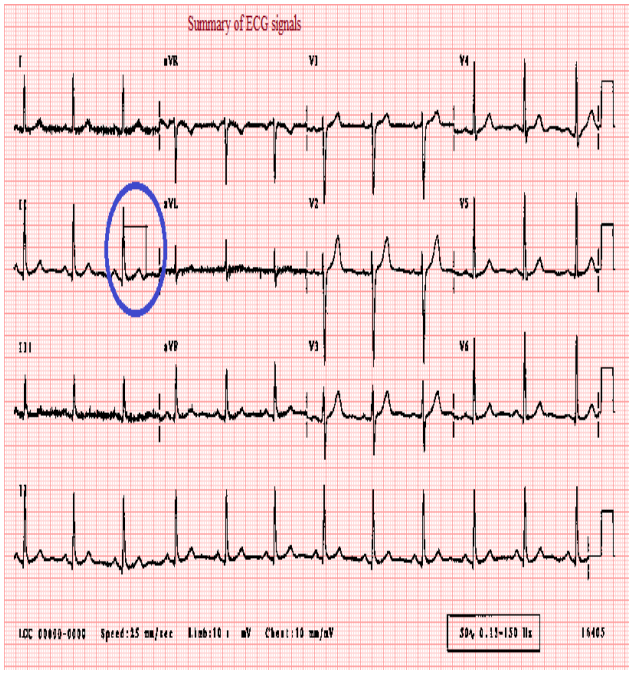


Fig. 7. Using the Caliper Tool to measure the duration of Heart Disease.

Its user interface additionally includes functionalities that resemble a conventional style in aiding physicians in their analysis. Fig. 6 shows how the information was displayed in a real-world sheet style. Caliper capability is incorporated on a GUI webpage to give a straightforward technique of measuring periods to manually evaluate through physicians. Another purple arrow from Fig. 7 and 8 represents the separation measurements findings from two separate places over each transmission for CKD and heart disease.

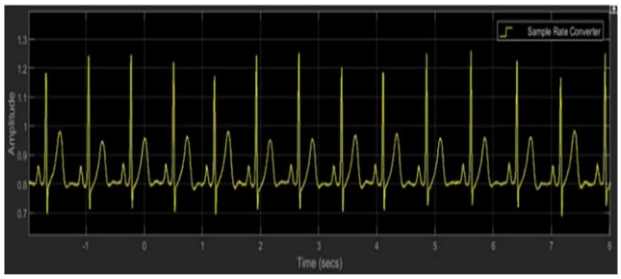


Fig. 8. Using the Caliper Tool to measure duration CKD.

V. Conclusion

The QT interval was a crucial statistic that may be used as a reliable indicator for cardiac illness. However, owing to the characteristics of ECG transmitter constituent waveforms, its nature of assessments, its absence of standards, & its huge quantity of information, this was difficult to evaluate autonomously. During beat-by-beat QT assessment, researchers proposed a novel method. Researchers also used the proposed technique to create an internet solution for QT assessment. In addition, the software included several essential features for instructions and automated QT evaluation. Researchers used the DICOM recorded standard to ensure interoperability that using existing recording equipment. To determine the heart rate, the attenuated ECG signals were analyzed for CKD and heart disease through Q wave spectral analysis. The displayed heart rate was well below the heart rate range of 79–82 beats per minute (bpm). The Q wave found was also similar to those of the QT algorithm. The filter ECG contained no noise or abnormalities in muscle motion, respiratory variability, or baseline wandering, and was within the expected peak amplitude range. There is no noise in the Q waveform detected; therefore the heart rate is correctly detected. T-wave has been filtered from 0.05 to 0.07 mV as expected. The heart rate of the analyzed signal ranges between 79 and 82 BPM. Consequently, the validity of the algorithm used to build simulation blocks for cardiac surveillance has been proven.

References

[1] Malek, A. S., Elnahrawy, A., Anwar, H., & Naeem, M. (2020). Automated detection of premature ventricular contraction in ECG signals using enhanced template matching algorithm. Biomedical Physics & Engineering Express, 6(1), 015024.

[2] Rueda, C., Larriba, Y., & Lamela, A. (2021). The hidden waves in the ECG uncovered revealing a sound automated interpretation method. Scientific reports, 11(1), 1-11.

[3] Chatterjee, S., Thakur, R. S., Yadav, R. N., Gupta, L., & Raghuvanshi, D. K. (2020). Review of noise removal techniques in ECG signals. IET Signal Processing, 14(9), 569-590.

[4] Fotiadou, E., Konopczyński, T., Hesser, J., &Vullings, R. (2020). End-to-end trained encoder-decoder convolutional neural network for fetal electrocardiogram signal denoising. Physiological measurement, 41(1), 015005.

[5] Xue, J., & Yu, L. (2021). Applications of Machine Learning in Ambulatory ECG. Hearts, 2(4), 472-494.

[6] Etzkorn, L. H., Heravi, A. S., Wu, K. C., Post, W. S., Urbanek, J., &Crainiceanu, C. (2021). Classification of Free-Living Body Posture with ECG Patch Accelerometers: Application to the Multicenter AIDS Cohort Study. bioRxiv.

[7] Atanasov, V., Sivkov, Y., &Velikov, N. (2020, September). An approach of Feature extraction of ECG signal of CLAS database. In 2020 International Conference on Biomedical Innovations and Applications (BIA) (pp. 93-96). IEEE.

[8] Latchoumi, T. P., Balamurugan, K., Dinesh, K., & Ezhilarasi, T. P. (2019). Particle swarm optimization approach for waterjet cavitation peening. Measurement, 141, 184-189.

[9] Zhang, S., Zhao, T., Peng, C., Li, Q., & Zhang, X. (2020, June). Design and Implementation of A Novel Real-Time P-QRS-T Waves Detection Algorithm. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 1, pp. 1109-1112). IEEE.

[10] Kashou, A. H., Ko, W. Y., Attia, Z. I., Cohen, M. S., Friedman, P. A., & Noseworthy, P. A. (2020). A comprehensive artificial intelligence-enabled electrocardiogram interpretation program. Cardiovascular Digital Health Journal, 1(2), 62-70.

[11] Pilia, N., Nagel, C., Lenis, G., Becker, S., Dössel, O., & Loewe, A. (2021). ECGdeli-an open source ecg delineation toolbox for MATLAB. SoftwareX, 13, 100639.

[12] Zhang, Q., & Li, X. (2021, April). Analysis and Application of Bayesian Network and Qt View Framework in Network Fault Location. In 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 981-984). IEEE.

[13] Garikapati, P., Balamurugan, K., Latchoumi, T. P., &Malkapuram, R. (2021). A Cluster-Profile Comparative Study on Machining AlSi 7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means. Silicon, 13, 961-972.

[14] Wang, Y., Wu, L., Chen, C., Jin, Z., Li, Z., & Wang, Y. (2020, August). P wave Detection in Electrocardiogram Based on Wavelet Transform and Differential Correction. In 2020 12th International Conference on Advanced Computational Intelligence (ICACI) (pp. 277-282). IEEE.

[15] Sun, J. Y., Shen, H., Qu, Q., Sun, W., & Kong, X. Q. (2021). The application of deep learning in electrocardiogram: Where we came from and where we should go?. International journal of cardiology.

[16] Ezhilarasi, T. P., Dilip, G., Latchoumi, T. P., & Balamurugan, K. (2020). UIP—A Smart Web Application to Manage Network Environments. In Proceedings of the Third International Conference on Computational Intelligence and Informatics (pp. 97-108). Springer, Singapore.

[17] Ohmuta, T., Mitsui, K., & Shibata, N. (2020). ECG QT-Interval measurement using wavelet transformation. Sensors, 20(16), 4578.

[18] Attia, Z. I., Harmon, D. M., Behr, E. R., & Friedman, P. A. (2021). Application of artificial intelligence to the electrocardiogram. European Heart Journal, 42(46), 4717-4730.

[19] Surekha, K. S., & Patil, B. P. (2021). Power line interference removal from electrocardiogram signal using multi-order adaptive LMS filtering. International Journal of Biomedical Engineering and Technology, 35(2), 135-151.

[20] Andršová, I., Hnatkova, K., Šišáková, M., Toman, O., Smetana, P., Huster, K. M., ... & Malik, M. (2021). Influence of heart rate correction formulas on QTc interval stability. Scientific Reports, 11(1), 1-21.

[21] Kumar, A., Ranganatham, R., Singh, S., Komaragiri, R., & Kumar, M. (2021). A robust digital ECG signal watermarking and compression using biorthogonal wavelet transform. Research on Biomedical Engineering, 37(1), 79-85.

[22] Guven, G., Guz, U., & Gürkan, H. (2022). A novel biometric identification system based on fingertip electrocardiogram and speech signals. Digital Signal Processing, 121, 103306.

Game-based Learning Increase Japanese Language Learning through Video Game

Yogi Udjaja, Puti Andam Suri, Ricky Satria Gunawan, Felix Hartanto
Computer Science Department, School of Computer Science
Bina Nusantara University, Jakarta
Indonesia 11480

***Abstract*—This research was purposed to test the effectiveness of learning the Japanese language through a video game. the video game is built for Personal Computer (PC) users to provide Japanese Language education through video games for teens and adults. The research methods used include literature studies of various books, journals, websites, and theories that can support the writing as well as defining the questions for questionnaires to collect useful data. The development of game application methodology used is Game-Based Learning with Enhanced Learning Experience and Knowledge Transfer (ELEKTRA) methodology, which consists of in-depth analysis of the target audience and learning materials. The effectiveness of video games are evaluated using pre-test and post-test methods. From this researches can be seen that video games are effective to increase the users' knowledge of the Japanese language. also, a video game has the capability to increase the user's interest in learning Japanese because of the visual form of the learning process that leads the user to stay engaged with the learning process.**

Keywords*—Video games; ELEKTRA; games-based learning; Japanese; JLPT N5

I. INTRODUCTION

From the previous research [1], [2], in 2018, Japan was the second most popular country for continuing education in the world after the UK. Even now, Japan is still in second place [1]. With so many students wanting to continue their education in Japan. In fact, that most Japanese people don't want to use a language other than the Japanese language. Hence, knowledge of the Japanese language is very important. From 100 students studying in Japan, 81% say that Japanese is difficult to understand [2].

To test fluency in Japanese, we need to take the Japanese Language Proficiency Test (JLPT). According to Haristiani & Firmansyah, JLPT is a well-known form of Japanese language proficiency test, or in Japanese, it is called Noryoku Shiken [3]. Noryoku Shiken has a Scale from level N5, which is the most basic level, to N1, which is the highest level. From all the tests carried out, Kanji is the most difficult test because Kanji should be integrated into the sentences in the exam, so mastering Kanji is mandatory.

Based on the sample vacancies from Tomodachi-Indo, the JLPT certificate is used as a condition for employment, and vacancies are also available in Indonesia. By understanding Japanese in depth, someone who has a certificate of Japanese language proficiency at the N1 level can get a salary of around 20 million. while for JLPT N2 certificate holders can get a salary of around 10 million, and even N3 and N4 can get vacancies if they can speak Japanese fluently [4].

In Indonesia, the biggest problems in Japanese language education according to Mayantara in 2012 were "Inadequate facilities/equipment" (57.1%), "Lack of information on teaching methods" (51.9%), "Lack of teaching materials" (44.4%), and "Lack of information about Japanese culture and society (39.1%). This causes the problem of lack of intention to learn (33.1%) [5].

Therefore, the creation of a new and fun learning model should be made to increase knowledge of the Japanese language. Usually, the learning model used to learn Japanese is in the form of a textbook. However, often the language conveyed is difficult to understand and tends to be monotonous. But by using multimedia elements, students will be "tricked" by the effects provided such as graphics, animations and unconsciously also learn Japanese.

Currently, there are many approaches to learning through video games. based on research conducted by A. Alamri [6] Learning through video games has advantages including:

1) Increase motivation in learning.

2) Learning through video games is a safe learning model because students do not need to do outdoor activities but still have the same opportunity to apply their knowledge without wasting time and money.

3) Learning through video games is a technology that can be accessed anywhere and is friendly to students with disabilities.

For these reasons, a special video game for learning Japanese was made using the ELEKTRA methodology [7], ELEKTRA is a special method for game-based learning that contributes to the need for a multimedia-based Japanese learning methodology. This video game is made with standards and target abilities ranging from Hiragana, Katakana, to basic Kanji which are likely to appear in JLPT N5, for students who want to continue their education in Japan or just learn Japanese.

In the end, the study will measure the effectiveness of learning Japanese using video games with pre-test and post-test methods to be able to see an increase in someone's knowledge in learning Japanese.

II. Related Research

A. Japanese Language

The Japanese language is the national language used in Japan. There are three writing systems in the Japanese writing system, namely: Kana, Kanji and Romaji [8]. Kana is one of the Japanese syllabaries which is a phonetic symbol of Chinese characters and in the early ninth century was used by Japanese people to read Chinese characters [9], [10] and is divided into two models, namely, hiragana and katakana. Here hiragana is usually used for grammar instructions while katakana is usually used for words in a foreign language, Fig. 1.

Meanwhile, kanji are Japanese characters that are ideographic symbols derived from Chinese characters [11], [12] which indicate traditional nouns, adjectives, and verbs. Another Japanese writing system is Romaji. Romaji writing is currently extending not only for stylistic purposes but also to help people who do not speak Japanese to be able to imitate Japanese sounds.

Being able to memorize hiragana, katakana, and kanji characters is a must to be able to understand the Japanese Language because Japanese letters have a different model than the general alphabet [13]. So many researchers are trying to compare various ways so that someone can learn Japanese better and faster [1].

B. Typing Game

Since the presence of pong as a computerized table tennis game in 1972, the game industry has continued to develop into a very large entertainment business. Currently, video games continue to grow and are divided into many categories such as puzzles, strategy, simulations, RPGs, sports, etc. [14]. This is supported by a shift in the benefits of games which are not only entertaining but are also widely used as learning media [6].

In language learning, people often feel stressed, and it is very difficult to learn, especially when they have to practice the language in the real world. By utilizing this game it can reduce the pressure for someone while studying [14]. In addition, games can form a learner-centered environment [15]. With the right design and the right implementation process, games can increase motivation in learning, especially language learning which will form closer engagement if you play more often and engage in communication.

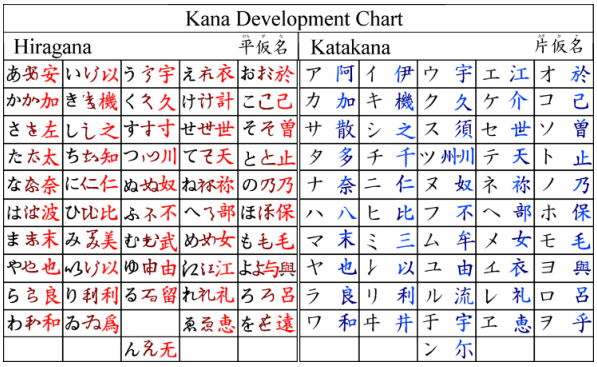


Fig. 1. Kana Development Chart [9].

Among the many types of games, typing games are the type of games that are very suitable for language learning. Where typing games can improve one's ability in typing can also train one's ability to spell and train the sentence structure (grammar) of a language of course in a more fun way [14]. Because of the benefits of typing games for language learning, now many typing games are available to learn various types of languages such as word games for folk play learning [16] and the kana no shensi for learning Japanese hiragana and katakana letters [17].

C. Motivation Through Game

Video games give rise to many types of motivation, starting from the desire to compete, fantasy, excitement, relaxation and others. This is not necessarily obtainable in traditional learning methods [18]. Because video games can evoke strong emotional feelings and provide a sense of satisfaction in the learning process [19]. If someone who is learning feels a pleasant experience while learning, then this will make that person motivated to continue learning [6].

Game-Based Learning Approach is highly recommended as a learning method because of its potential to practice decision-making skills and is suitable for use in various study areas [20], [21].

One of the studies conducted by Monter M et al. where he tested the process of learning English for students who used Arabic as their daily language. In his research he found that the use of video game in English class made students happier and some of the students continued the games in their spare time at home. Even further, in this research he suggested making the use of game applications as part of the curriculum and using the application. used throughout the semester to see the full potential of the use of game applications in the learning process [22].

III. Research Methods

The research method used is Game-Based Learning. According to Linek et al. Game-Based Learning is edutainment that utilizes the motivational and immersive aspects of video games in an educational context. The available methodology for Game-Based Learning is in the form of Enhanced Learning Experience and Knowledge Transfer (ELEKTRA) [7]. ELEKTRA has eight phases to ensure a clear workflow, which is shown in Fig. 2.

1) Phase 1: Identify instructional goals: At this stage, a search and analysis of the problem are carried out, and then the solution is sought by designing the targets and objectives of the learning video game. In the first phase, the target audience and how the game is implemented is determined as the next design determination. The target audience is then determined in phase 3. Requirement Analysis from Black Box Testing is also held in this phase.

2) Phase 2: Instructional analysis: In this phase, the team collects data that provides a solution to achieve the goals that have been determined from Phase 1. In the case of learning video games, this can be done by researching learning

materials, then, how to implement these learning materials as games.

3) Phase 3: Analyze learners and context of learning: In this phase, the focus is given to the analysis of the players themselves. The target audience was analyzed through a questionnaire survey. The results of the survey and questionnaire are used in this phase as a determination of entry skills for various levels in this game. For players, the entry skills for this game are as follows in brief:

- Hiragana: Understand how to write Hiragana
- Katakana: Understand how to write Katakana
- N5 Kanji: Understand how to write Hiragana and learn Kanji from outside sources.

Here also uses a questionnaire to determine the wishes of the players, and the results of the questionnaire are used as material for discussion in this phase.

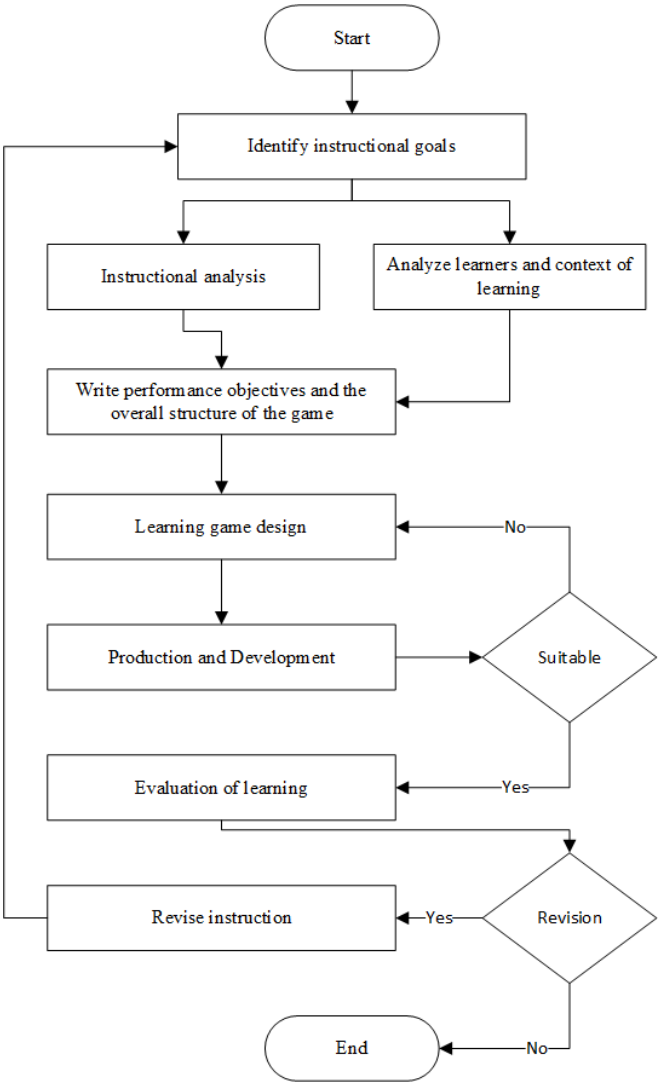


Fig. 2. ELEKTRA.

4) Phase 4: Write performance objectives and the overall structure of the game: From the results of Phase 3 and its entry-level, this information is used to determine the desired performance goals for players. After the entry level and performance targets are determined, then the problems and solutions are formulated. The overall structure of the game is formally explained using Storyboarding. Design Specifications are also determined in this phase to provide clear objectives to the programmer.

5) Phase 5: Learning game design: After finishing designing targets based on the results of the previous stages, the design of the game begins with Game Design Document, display design, game performance design, and test design. This stage is also the first stage of the research practicum from the previous stages, and over time, features that are not feasible or cannot be implemented are discarded, or added to create the most appropriate User Experience for the target audience. The way this game works is as follows:

- Players start at the Main Menu and can choose the type of Japanese writing, namely Hiragana, Katakana, and JLPT N5.
- After selecting a level, Letters/words will appear on the screen, then the player is expected to type the spelling before the letters/words fall to the bottom, taking one of the player's three lives.
- The game ends when lives reach empty. Then, the player can choose to restart again or return to the home screen.

6) Phase 6: Production and Development: After the design process has been completed, the programming of this game begins using the C# language in Unity. At this stage, everything that has been learned from the previous stage is used to design a game according to the design and especially the targets that have been determined from the first stage.

This phase also includes the testing phase. That is internally and externally. Internal testing or referred to as the alpha version is carried out internally to test every part of the game to look for bugs that can appear while playing, find the root of the problem, and fix the game using Black Box Testing theory.

After the game has been cleaned of as many bugs and features as possible are added and removed through game design changes (phase 5), the game is entered into the External testing section, or known as the beta version, where the game is given to questionnaire participants and managed through a pre-test and post-test. Constraints found from the External testing phase can be managed and updated to a newer beta version.

The process will continue to return to phase 5 every time there is an update until no more changes need to be made. If no changes are needed, the development process continues to phase 7.

7) Phase 7: Evaluation of learning: Evaluation of the learning process is carried out by using the results of the post-test of the questionnaire participants as evidence and results of the game design that has been made, and then determining whether the game has achieved the desired target.

8) Phase 8: Revise instruction: The revision of the Evaluation results is used as a guide for features that should be added, changes that need to be made before the game is finally ready to be released and used as determined from Phase 1.

IV. Result and Analysis

Fig. 3 to 10 are the results of Japanese language video game-based learning called "Typing Japan". This video game will run on a PC/Desktop with a minimum operating system of Windows 7 SP1+, macOS10.12+ or Ubuntu 16.04+, and must-have graphics capabilities to run DirectX 10.

This video game is divided into two main menus in this game, which are as follows:

1) Writing Style, wherein this menu the user will determine the type of word to be played. In this case, there are 3 types of word choices, namely: Hiragana, katakana, and N5 Kanji. and in the process of selecting the type of word to be played, the user will also be given the option to choose the level they want to learn, for example on the Hiragana menu there will be options for Syllables I, Syllables II, Dakuten, Yoon, All Syllables, and Full Set.

2) The game itself: The way to play this game application is by typing the spelling of the letter from the letters that fall before touching the bottom of the screen. When the player starts writing from the letter, the letter changes color to red. When the player finishes writing the letter, the letter turns green and disappears. In this game, the player is initially given three lives. When the letter hits the bottom of the screen, the player's life will be reduced by one. The game ends when the player has run out of lives. Click the back button to return to the Writing Style Mode menu.

This game application was evaluated to 30 respondents who wanted to learn Japanese with several methods:

1) Obtaining feedback from the target user through a questionnaire to determine how effectively the game “Typing Japan” helps increase insight by collecting comparative data from the pre-play and post-play Questionnaire. Collect the requirements needed by the user to be implemented into game applications.

2) The Alpha Testing process uses the Black Box Testing methodology to check if the video game works well.

3) Getting feedback from Open Beta players will help evaluate the game from the player's perspective to maximize Game Experience (GX) when the game is released and published to the public.

At the testing stage, pretest and posttest were carried out to determine the increase in knowledge after using the application, so that an increase in the ability of 27 of 30 users is obtained.

Fig. 3. Main Menu.

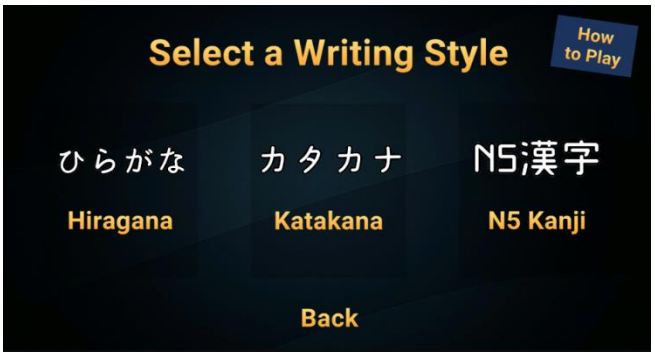


Fig. 4. Writing Style Mode.

Fig. 5. How to Play.

Fig. 6. Hiragana Menu.

Fig. 7. Katakana Menu.

Fig. 8. N5 Kanji Menu.

Fig. 9. Gameplay.

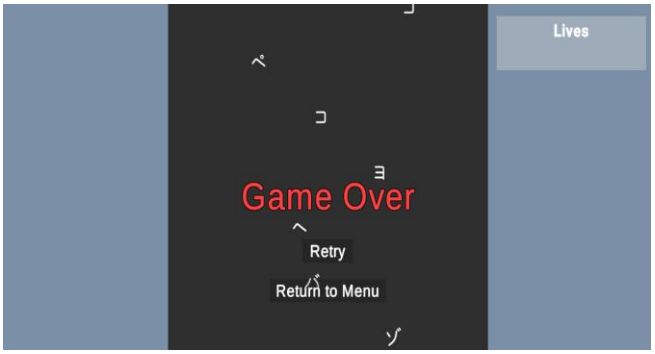


Fig. 10. Game Over.

Besides that, the use of video games can increase a person's interest in continuing to learn Japanese. 23 out of 30 users stated that they are very happy and interested in continuing to learn. While 7 other users stated that they were quite happy with the learning process through video games.

In addition to conducting user satisfaction questionnaires, Alpha Testing was also carried out. The entire Alpha Testing process is carried out internally with the Black Box Testing methodology. Black-box testing is testing done without knowing the source code of the project. Based on the test results, all functions of the application can run well and can be accepted by the user.

In terms of appearance and experience in playing, based on the results of the post-playing questionnaire, it is stated that the game still needs a lot of improvement. This can be in the form of sound effects, music, feedback when writing a letter/word incorrectly, as well as other feedback such as scoring, grading, and sense of progress. All of this requires more experience in database programming. In addition, the UI on the Main Menu and Gameplay has been stated quite clearly and the gameplay can be stated smoothly.

V. Discussion and Conclusion

Everyone has different abilities. There are people who can quickly understand something with just one look, while there are people who have to study hard to understand something. In terms of language learning, the intensity of a person to keep practicing foreign language skills will greatly affect a person's learning speed. The more often a person trains his abilities, the more a person's ability to speak a foreign language will increase.

The existence of video games can be a method of training language skills that attract users' interest and support users to practice their language skills on an ongoing basis because the display is much more attractive than a textbook can also be accessed when outside the classroom.

This learning process will be even more interesting if it is supported by good audiovisuals and sound effects. It will be even more interesting if the video games are supported by technology that utilizes the internet such as multiplayer, leaderboard, etc. [23]. These things can speed up the learning process by taking advantage of the competitiveness and resilience of the user.

References

[1] Y. Udjaja, "Gamification Assisted Language Learning for Japanese Language Using Expert Point Cloud Recognizer," Int. J. Comput. Games Technol., vol. 2018, 2018, doi: 10.1155/2018/9085179.

[2] Y. Udjaja, Renaldi, Steven, K. Tanuwijaya, and I. K. Wairooy, "The use of role playing game for Japanese language learning," Procedia Comput. Sci., vol. 157, pp. 298–305, 2019, doi: 10.1016/j.procs.2019.08.170.

[3] N. Haristiani and D. B. Firmansyah, "Android application for enhancing Japanese JLPT N5 kanji ability," J. Eng. Sci. Technol., vol. 12, no. Special Issue 10, pp. 106–114, 2017.

[4] "Contoh Lowongan Kerja untuk Pemegang Sertifikat JLPT | Tomodachi | Kursus Bahasa Jepang." https://www.tomodachi-indo.com/lowongan-kerja/ (accessed Dec. 11, 2021).

[5] "Survei Lembaga Pendidikan Bahasa Jepang di Indonesia tahun 2012 – Mayantara School." https://mayantara.sch.id/artikel/survei-lembaga-pendidikan-bahasa-jepang-di-indonesia-tahun-2012.htm (accessed Dec. 11, 2021).

[6] A. Alamri, "Should Video Games Be Included in the Learning Process?," Int. J. Educ., vol. 8, no. 1, p. 23, 2016, doi: 10.5296/ije.v8i1.8388.

[7] S. B. Linek, D. Schwarz, M. Bopp, and D. Albert, "Game-based learning:Conceptual methodology for creating educational games," WEBIST 2009 - Proc. 5th Int. Conf. Web Inf. Syst. Technol., no.

January, pp. 135–142, 2009.

[8] H. Rose, “Unique challenges of learning to write in the Japanese writing system,” L2 Writ. Beyond English, no. May, pp. 78–94, 2019, doi: 10.21832/9781788923132-008.

[9] K. Nute, “Toward a Test of Cultural Misappropriation,” no. April, 2020, doi: 10.18848/2327-0055/CGP/v17i02.

[10] NHK, “Hiragana | Cara Mudah Berbahasa Jepang | NHK WORLD-JAPAN.” https://www.nhk.or.jp/lesson/id/letters/hiragana.html (accessed Dec. 11, 2021).

[11] Y. Okuyama, “CALL Vocabulary Learning in Japanese: Does Romaji Help Beginners Learn More Words?,” CALICO J., vol. 24, no. 2, pp. 355–379, 2013, doi: 10.1558/cj.v24i2.355-379.

[12] A. S. Dylman and M. Kikutani, “The role of semantic processing in reading Japanese orthographies: an investigation using a script-switch paradigm,” Read. Writ., vol. 31, no. 3, pp. 503–531, 2018, doi: 10.1007/s11145-017-9796-3.

[13] T. Ogino, K. Hanafusa, T. Morooka, A. Takeuchi, M. Oka, and Y. Ohtsuka, “Predicting the reading skill of Japanese children,” Brain Dev., vol. 39, no. 2, pp. 112–121, 2017, doi: 10.1016/j.braindev.2016.08.006.

[14] T. L. Wang, T. K. Chen, and Y. F. Tseng, “An leaner-centred, game-based, learning framework for typing games in english course,” 3CA 2010 - 2010 Int. Symp. Comput. Commun. Control Autom., vol. 1, no. June 2010, pp. 93–95, 2010, doi: 10.1109/3CA.2010.5533723.

[15] B. Klimova and J. Kacet, “Efficacy of computer games on language learning,” Turkish Online J. Educ. Technol., vol. 16, no. 4, pp. 19–26, 2017.

[16] Y. Bae, H. Choe, T. Lee, and T. Kim, “Folk Play Learning System Based on Word Games,” pp. 0–1, 2004.

[17] K. Stubbs, “Kana no Senshi (Kana Warrior): A new interface for learning Japanese characters,” Conf. Hum. Factors Comput. Syst. - Proc., no. January 2003, pp. 894–895, 2003, doi: 10.1145/765891.766054.

[18] Y. Udjaja, V. S. Guizot, & N. Chandra, "Gamification for elementary mathematics learning in Indonesia." International Journal of Electrical and Computer Engineering (IJECE), 8(6), 2018.

[19] Y. Udjaja, Sasmoko, A. S. Rumapea, F. A. Putra, & T. Rahmansyah, "Architecture of High-Order Thinking Skills Game to Improve Ability." In 2019 IEEE International Conference on Engineering, Technology and Education (TALE) (pp. 1-4). IEEE, 2019, December.

[20] P. Kinnerk, S. Harvey, C. MacDonncha, and M. Lyons, “A Review of the Game-Based Approaches to Coaching Literature in Competitive Team Sport Settings,” Quest, vol. 70, no. 4, pp. 401–418, 2018, doi: 10.1080/00336297.2018.1439390.

[21] F. S. Breien and B. Wasson, “Narrative categorization in digital game-based learning: Engagement, motivation & learning,” Br. J. Educ. Technol., vol. 52, pp. 2021–91, 2020, doi: 10.1111/bjet.13004.

[22] M. M. Elaish, N. A. Ghani, L. Shuib, and A. Al-Haiqi, “Development of a Mobile Game Application to Boost Students’ Motivation in Learning English Vocabulary,” IEEE Access, vol. 7, pp. 13326–13337, 2019, doi: 10.1109/ACCESS.2019.2891504.

[23] Byun, JaeHwan & Loh, Christian. (2015). Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments. Computers in Human Behavior. 10.1016/j.chb.2014.12.052.

Fuzzy-set Theory to Support the Design of an Augmentative and Alternative Communication Systems for Aphasia Individuals

Md. Sazzad Hossain
Department of Computer Science and Engineering
Mawlana Bhashani Science and Technology University
Tangail, Bangladesh

***Abstract*—This paper presents a new design of an Augmentative and Alternative Communication (AAC) systems for conveying delicate feelings or emotions of aphasia individuals, which is based on Fuzzy-set theory. Fuzzy-set theory is crucial in addressing the ambiguity of linguistic terms used and judgments made by aphasia individuals. Due to the communication difficulties of aphasia individuals, their insights were assigned in triangular fuzzy membership functions during the design process of AAC systems. In the proposed design of AAC systems, the delicate feelings or emotions were expressed as a scale, and candidate(s) of delicate feelings or emotions were shown based on their specified position. If the candidate(s) cannot properly convey the desired delicate feelings or emotions, then the corresponding fuzzy membership function can be realized by controlling its position. The proposed method has the advantage of being able to be conveyed the exact want and needs of delicate feelings or emotions during communication. Experimental result shows that conveying delicate feelings or emotions of the aphasia individual could be improved by 50 percent using the proposed design of AAC systems.**

***Keywords*—Aphasia; augmentative and alternative communication; human factors; fuzzy-set theory**

I. Introduction

Aphasia is a speech and language impairment, caused by acquired brain damage [1]. A stroke is the most common cause of aphasia. It also caused by other types of acquired brain injuries include traumatic brain injuries, brain tumors, and anoxia. The incidence of aphasia after stroke is about 20% to 38% in the acute phase [2][3]. Moreover, brain damage often causes hemiparesis, which is a weakness or inability to move on one side of the body. Basically, right-sided hemiparesis involves injury to the left side of the brain, which controls language and speaking. People with this type of hemiparesis may have trouble speaking and/or understanding what people are saying.

Depending on the specific locations of brain damage, the severity and pattern of aphasic symptoms vary from person to person. Broca's type aphasia is called non-fluent aphasia. They have difficulty in speaking. Wernicke's type aphasia is called fluent aphasia. They have difficulty in understanding. However, speech-language therapist (SLT) supports aphasia individuals to perform their daily activities.

Aphasia affects individuals' ability to speak, to understand speech, to read and to write. These language difficulties seriously hamper their daily communications [4]. It is noted that other disabilities caused by brain damage such as motor speech disorder (e.g., dysarthria, dysphonia or apraxia of speech) affect intellectual capabilities. People with these disabilities have no difficulties in finding the words they wish to say, and they report no difficulties with reading, writing, or auditory comprehension. On the other hand, aphasia individuals cannot communicate properly by their own words because their brain areas are restricted to process primarily speech and language, but their intelligence is intact. Thus, communication barriers often stigmatized disabled people and can further exacerbate the difficulties in quality of life (QOL) [1]. Consequently, aphasia individuals live at home with their families after they leave the hospital. During this time, a speech-language therapist (SLT) supports the aphasia individuals to improve their communication. SLT supports aphasia individuals based on the diverse symptoms such as non-fluent, fluent, mild, moderate, severe, unable to read, unable to write or unable to understand. SLT supports them in various ways of alternative communication such as memo writing, using picture boards, picture cards or gestures, which are examples of AAC systems. Although these types of support of SLT through rehabilitation improve the communication skills of aphasia individuals, they still cannot convey their delicate feelings or emotions to others by the current AAC systems.

Basically, AAC systems include pictures and symbols, which use electronic devices to adapt voice output communication aids, methods, and techniques [4]. In this regard, several computer applications and many portable devices are available for such communication support. As a result, AAC system is now widely used on available mobile devices, tablets, and PCs. [5]. For example, Proloquo2Go is a common AAC system based on a folder structure, each displaying a box with images and symbols, or providing a text typing input box [6]. A selected symbol can be spoken with natural sounds or via a machine voice. The main problem of the existing AAC systems is the difficulty expressing the exact wants or needs by the aphasia individuals with their delicate feelings or emotions. The reason is that the design process of popular design approaches cannot include the insights of aphasia individuals in the AAC systems because of their

communication difficulties. To solve this problem, Fuzzy-set theory can support the design of the AAC systems to include the insights of aphasia individuals to convey their delicate feelings or emotions.

The popular design approaches to design AAC systems for disabled people are Barrier-free Design, Universal Design, Design for All or Inclusive Design. Barrier-free Design is specially introduced to remove architectural obstacles for disabled people [7][8]. On the other hand, Universal Design and Design for All look for a design solution that can support everyone including people with disabilities [9][10]. However, Universal Design is insufficient to cover everyone's needs. First, designers acquire needs of product or environment from different user groups including those with disabilities. Then designers identify common needs that can support all user groups. Designers think that design solutions are enough to fulfill these common needs. Finally, designers complete a design with these common needs. As a result, disabled users' needs are partially included in the common needs of Universal Design solutions. Nevertheless, these common needs are not sufficient to fulfill disabled users' needs. Thus, disabled users cannot use the design solutions. It is noted that these disabled users are extreme users because they represent the extreme end of the usability spectrum and are most affected by poor design solutions as shown in Fig. 1. Extreme users may have exaggerated needs, thought or behavior compared to the typical users. In addition, extreme users can offer unique insights about the products and inspire a different way of thinking about current and future users. For this reason, Universal Design seems impractical and ineffective for extreme users with common views of the design [11].

On the other hand, Inclusive Design is an approach of designing with extreme users to find different ways for the access of products or environments. In Inclusive Design, insights of extreme users' need to be included in the products or environments. Therefore, this study used inclusive design to include the insights of aphasia individuals in the AAC systems. However, the insights of an aphasia individuals are not included properly in the AAC systems at the design process of Inclusive Design. Here, insights of aphasia individuals are delicate feelings or emotions. Although an aphasia individual can identify wants and needs from their experiences, they cannot express them to the designer. As a result, their exact wants and needs are not met by existing design. Only the aphasia individual can give insights into different ways of participation to access the AAC systems but other people like designer, SLT or proxies cannot include an aphasia individual's insights. Consequently, design cannot cover the expectation of aphasia individuals to access the AAC systems in different ways. Therefore, design process of AAC systems needs support to improve communication skills of aphasia individuals. To address the above-mentioned issues, this study used Fuzzy-set theory to support the design process of AAC systems for aphasia individuals regardless of their communication difficulties. In other words, Fuzzy-set theory includes the delicate feelings or emotions of aphasia individuals to support the design of AAC systems.

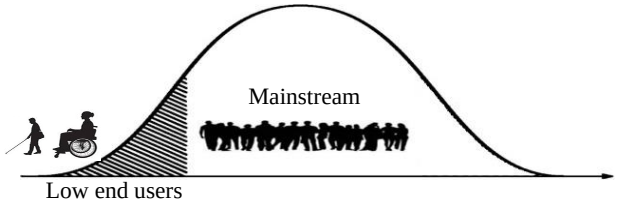


Fig. 1. Usability Spectrum of Product or Environments.

II. Related Work

The majority of previous studies used User-centered Design (UCD) or Participatory Design (PD) approaches to design an AAC systems [12][13][14]. Most UCD and PD methods make fundamental assumptions about the communication skills of those who will participate. They are founded on the premise that participants will have the requisite skills, for example, to communicate orally, to understand and produce written text, to comply with instructions. Those who do not have these skills cannot readily participate. Therefore, these AAC systems cannot fulfill the needs of aphasia individuals.

Several AAC systems have been developed using HCD or PD to assist people with aphasia, but a small number of AAC systems have been designed to assist communication for people with aphasia. Mahmud et al. implemented an email tool for language impairments using UCD approach [15]. Their tool was developed for sending email where other tasks cannot be performed. Thus, this tool is insufficient to assist other task for daily activities of diverse disabled users. Those that have used a PD approach to design have mostly used proxies. In other words, either SLT played the roles of the aphasic participant or the caregivers of aphasic participants provided feedback. For example, SLT played as proxies for aphasia users in the development of PhotoTalk [16], an application that allows people with aphasia to capture and manage digital photographs to support face-to-face communication. Thus, designers were able to specify only few real user requirements for diverse aphasia individuals. Koppenol et al. similarly designed an application that uses photographs to support communication and used SLT as proxies [17]. Kane et al. design a context-aware communication tool for improving interpersonal communication for people with aphasia [18]. They used PD at the design process, but they considered only two context such as current location and conversation partner. Hossain et al. also designed a context-aware communication tool to assist people with aphasia for improving their communication skills [19]. They used context history to get suggestion during communication. In the above-mentioned AAC systems, either the SLT played the roles of aphasic participants or the caregivers of them provided feedback to the designers. As a result, the design of AAC systems can partially fulfill the needs of aphasia individuals because they are implemented based on the common needs. Thus, many aphasia individuals still cannot covey their exact needs and wants with their delicate feelings or emotions using the existing AAC systems. The reason is that the insights of aphasia individuals did not consider at design process. In other words, aphasia individuals can only provide their insights in Fuzzy form which are not included by designers or SLT. For this reason, Fuzzy-set theory is crucial to

collect the insights of aphasia individuals from the real participants in the design of AAC systems. Therefore, the objective of this paper is to design an AAC system with the support of Fuzzy-set theory.

III. METHODS

The overall design process of AAC systems is shown in Fig. 2. The process starts from the identification of interaction problems of old design. Then the design solution is implemented using the insights of aphasia individuals for conveying delicate feelings or emotions. Finally, the design solution is evaluated by aphasia individuals.

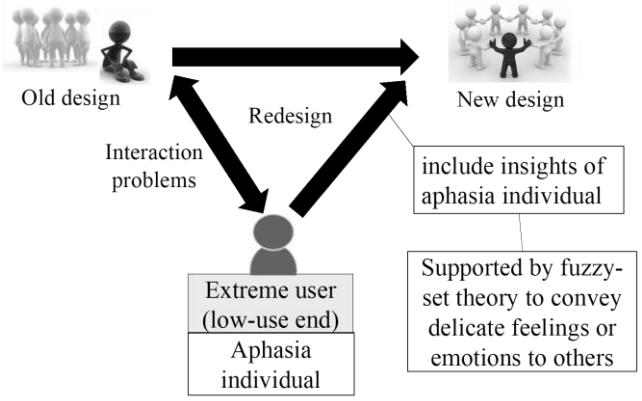


Fig. 2. The Design Process for AAC Systems for Aphasia Individuals.

A. Identification of Interaction Problems for Conveying Delicate Feelings or Emotions

The identification of interaction problems is completed by aphasia individuals as extreme users through a survey questionnaire. The survey questionnaire is offered to the aphasia individuals about delicate feelings or emotions. We would like to ask about 1) How much important is it to convey delicate feelings to others in order to actively participate in society? 2) How much can they convey their delicate feelings to others? and 3) How much can they understand the feelings of others? The interaction problems are identified based on the responses of each aphasia individual.

B. Redesign for Conveying Delicate Feelings or Emotions

The purpose of design is to include the insights of aphasia individuals for sufficient communication. In this study, the insights of aphasia individuals are delicate feelings or emotions. Aphasia individuals have difficulties to express their own thoughts verbally. Due to their difficulties, the delicate feelings or emotions are specified in fuzzy membership functions. Basically, they indicate delicate feelings or emotions on the scale of survey questionnaire during redesign. Their indicated positions for delicate feelings or emotions are not appropriate. Thus, a small amount of distance Δx needs to be managed so that fuzzy membership functions fit more based on their situations.

Aphasia individuals can convey their delicate feelings or emotions by pointing to a scale instead of their own voice. Basically, the scale represents the in-between situations (e.g. physical conditions, tiredness, etc.) of aphasia individuals. Thus, the delicate feelings or emotions were expressed after the redesign as a scale, and candidate(s) of them were shown based on his/her specified position as shown in Fig. 3. If the candidate(s) cannot properly convey their delicate feelings or emotions, then the corresponding function can be realized by controlling its position from x to (x+Δx) or (x-Δx) to adjust his/her delicate feelings or emotions.

$$\mu_{\tilde{M}}(x) = \begin{cases} 0 & x < \alpha \\ \frac{x-\alpha}{\beta-\alpha} & \alpha \le x \le \beta \\ \frac{\gamma-x}{\gamma-\beta} & \beta \le x \le \gamma \\ 0 & x > \gamma \end{cases} \quad (1)$$

To allocate the fuzzy membership functions for corresponding delicate feelings or emotions, they provide their insights during design as shown in Fig. 3. The membership functions of these fuzzy sets are denoted by equation (1), where (α, β, γ) denote the left-hand number, middle number, and right-hand number of each candidate of delicate feelings or emotions, respectively. For example, the triangular fuzzy number for 'very bad' can be represented by 0, 1, 1.5 as shown in Fig. 3. The triangular fuzzy membership numbers for delicate feelings or emotions are shown in Table I. The order of candidate(s) will be displayed according to the maximum possibility value of each candidate. The overall procedure for conveying delicate feelings or emotions based on the pointed location of aphasia individuals is represented in Fig. 4.

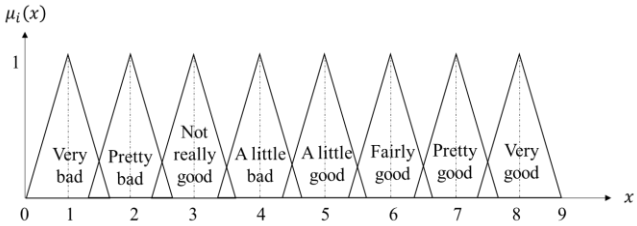


Fig. 3. Delicate Feelings or Emotions in Triangular Fuzzy Membership Function based on the Opinions of Aphasia Individuals.

TABLE I. TRIANGULAR FUZZY MEMBERSHIP NUMBERS FOR CONVEYING DELICATE FEELINGS OR EMOTIONS

Delicate feelings or emotions	Triangular fuzzy number
Very bad	(0, 1, 1.5)
Pretty bad	(1.25, 2, 2.5)
Not really good	(2.25, 3, 3.5)
A little bad	(3.25, 4, 4.5)
A little good	(4.25, 5, 5.5)
A fairly good	(5.25, 6, 6.5)
Pretty good	(6.25, 7, 7.5)
Very good	(7.25, 8, 9)

For example, an aphasia individual feels 'Not really good' and he wants to convey it to others. When aphasia individuals point an area between 'very bad' and 'very good' for physical condition, the possible candidates of delicate feeling are regarded as 'Pretty bad' and 'Not really good' based on the procedure as shown in Fig. 4. These candidates are displayed as shown in Fig. 5 from the Table II. Suppose the possibility

value for these two candidates 'Pretty bad' and 'Not really good' are 0.5 and 0.3, respectively. The candidate 'Pretty bad' holds the top position of the list. The aphasia individual realized that the candidate 'Not really good' is not on the top position. He/she moves the pointed location of fuzzy membership function from x to (x+Δx) or (x-Δx) to adjust his/her delicate feelings or emotions.

Step 1:	Identify I and $\mu_i(x), \forall i \in I = \{1, 2, \ldots, n\}$. Here, I is the index of delicate feelings or emotions $\mu_1(x), \mu_2(x), \mu_3(x), \ldots, \mu_n(x)$. n is the no. of index.
Step 2:	Ask a question to the aphasia individual about his/her current conditions.
Step 3:	Ask him/her to answer the question with a pointed location x.
Step 4:	For the pointed location x, calculate $\mu_i(x), \forall i \in I$. Identify the Set $I' = \{i \mid \mu_i(x) > 0\}$.
Step 5:	Show him/her the list of delicate feelings or emotions corresp onding to $i' \in I'$.
Step 6:	If the aphasia individual selects his/her desired delicate feelin gs or emotions as i'' index from I'. Then, go to Step 7. Other wise, go to Step 8.
Step 7:	For all x, set $\mu_{i''}(x) := \mu_{i''}(x + \Delta x)$ *or* $\mu_{i''}(x) := \mu_{i''}(x - \Delta x)$. Go to Step 10.
Step 8:	For all $i \in I'$, set $\mu_i(x) := \mu_i(x + \Delta x)$ *or* $\mu_i(x) := \mu_i(x - \Delta x)$. Go to Step 9.
Step 9:	Repeat steps 3 to 6 for new candidate list of delicate feelings or emotions.
Step 10:	End procedure.

Fig. 4. Procedure for Conveying Delicate Feelings or Emotions by Fuzzy-Set Theory.

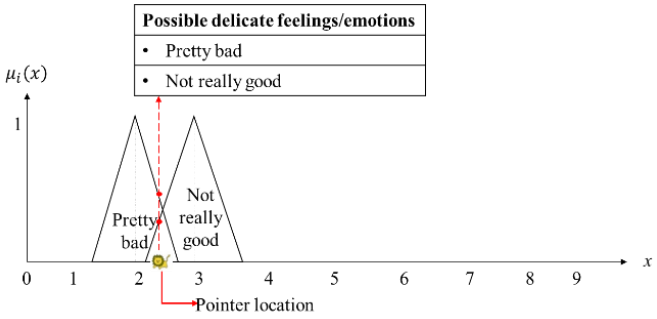


Fig. 5. Pointed Location and Corresponding Delicate Feelings or Emotions.

The possibility values of candidate are changed to 0.4 for 'Not really good' and 0.2 for 'Pretty bad' as shown in Fig. 6. Thus, the candidate 'Not really good' holds the top position of the candidate list.

C. Case Study to Support the Design an AAC Systems by the Fuzzy-set Theory

A case study was conducted through survey questionnaires with the participation of two aphasia individuals for conveying delicate feelings or emotions. A prototype for conveying physical conditions and tiredness was also provided with the survey questionnaires. The purpose of the first survey questionnaire was to identify the interaction problems for conveying delicate feelings or emotions. The second survey questionnaire was used to allocate the fuzzy membership functions for the corresponding delicate feelings or emotions for sufficient communication. Finally, the prototype was improved based on the interaction problems and the fuzzy membership functions through redesign.

TABLE II. DELICATE FEELINGS OR EMOTIONS TO CONVEY PHYSICAL CONDITIONS AND TIREDNESS

Feelings or emotions about	**Delicate feelings**	
	Worst to usual	**Usual to best**
Physical condition	Very bad	Very good
	Pretty bad	Pretty good
	Not really good	A fairly good
	A little bad	A little good
Tiredness	Quite tired	Not a little tired
	Very tired	Not tired at all
	Somewhat tired	Not so tired
	A little tired	Not too tired

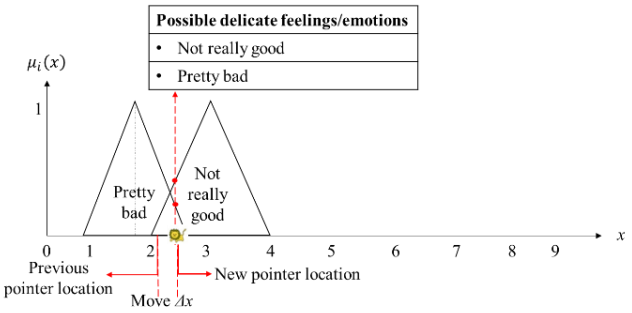


Fig. 6. After Moving the Pointer Δx, the New Pointed Location and Corresponding Delicate Feelings or Emotions.

TABLE III. RESPONSES OF SURVEY QUESTIONNAIRE FROM TWO APHASIA INDIVIDUALS

Questions	**Responses**	
	Aphasia individual 1	**Aphasia individual 2**
How much important is it to convey delicate feelings to others in order to actively participate in society?	Can do a little	Can do a little
How much can you convey the delicate feelings to others?	Can do a little	Can do a little
How much can you understand the delicate feelings of others?	Somewhat important	A little important
How much can you convey physical conditions to others?	Can do some extent	Can do some extent
How much can you understand physical conditions of others?	Cannot do	Cannot do
How much important to convey the physical conditions to actively participate in society?	Very important	A little important
How much can you convey tiredness to others?	Can do some extent	Cannot do
How much can you understand the tiredness of others?	Can do a little	Cannot do
How much important to convey the tiredness to actively participate in society?	Very important	Unimportant

D. Identification of Interaction Problems for Conveying Delicate Feelings or Emotions

To identify interaction problems, the responses of two aphasia individuals were collected as shown in Table III. From the two participants, the first aphasia individual is male, and he is 58 years old. He suffered from Broca's aphasia with the problem of memory impairment. He suffered from two strokes, one in March 2015 and another in November 2016. He can communicate with others, but his speech is non-fluent. He lives at home with his wife and children. The second aphasia individual is also male, he is 40 years old. He suffered from stroke at February 2018. He lives with his mother. He can communicate with others with a little fluent speech by rehabilitation. Thus, he returned his job. From the viewpoint of second aphasia individual, the provided prototype will be useful to more severe aphasia individuals than him.

The first aphasia individual responded that it was important in some extent for him to convey delicate feelings or emotions to others. He can communicate with others, but his speech is non-fluent. It was also very important for him to convey his physical conditions and tiredness to others in order to actively participate in society. He could convey his physical conditions and tiredness in 'some extent' to others. He could understand the tiredness 'a little', but he could not understand physical conditions of others. He also has memory impairment with Broca aphasia. He needs supports to convey his physical conditions and tiredness to others. Nevertheless, he cannot recall sometimes any words in his communication due to memory impairment and symptoms of Broca aphasia. As a result, he cannot express his delicate feelings properly by his own voice. He has recovered his communication abilities by his efforts and supports from SLT and his family. He wants to convey his physical conditions and tiredness more properly to others with his delicate feelings or emotions. The second aphasia individual was milder than the first one. He can communicate with others with a little fluent speech, and he has returned to his job by rehabilitation. Therefore, it was less important for him to convey his physical conditions and tiredness to others and to understand physical conditions and tiredness of others. However, he responded that his delicate feelings or emotions could be conveyed to others in his daily life. He also wants to convey more of his delicate feelings or emotions to others by the new design of AAC systems.

E. Allocation of the Fuzzy Membership Functions to Convey Delicate Feelings or Emotions

The first aphasia individual did not find the proper words to communicate with others based on his pointed location on the scale before the design. His pointed locations were allocated to the triangular fuzzy membership functions using a survey questionnaire during the design in order to find the proper delicate feelings or emotions when communicating with others. However, his specific position of fuzzy membership functions for corresponding delicate feelings or emotions was not proper. Due to the slope of the triangular fuzzy membership functions, it is possible to change the order of his delicate feelings or emotions based on his pointed location. The change of order of delicate feelings or emotions is done by controlling Δx on the scale. Thus, fuzzy membership functions fit more based on his situation to find proper delicate feelings or emotions during communication. According to his pointed location, the fuzzy membership functions for physical conditions and tiredness are shown in Fig. 7 and Fig. 8.

He pointed to the delicate feelings or emotions 'very bad' in the first position between bad and usual physical condition. He thought the delicate feelings or emotions 'not really good' and 'pretty bad' in a very close location on the analog scale. The candidates 'not really good' and 'pretty bad' take places respectively after the 'very bad' position. He also thought that 'a little bad' is close to usual condition and thus he pointed it as a neighbor of usual. He thought that 'pretty good' and 'very good' are close and he pointed 'pretty good' and 'very good' respectively on the last two positions. Finally, he pointed the candidates 'fairly good' and 'a little good' after the usual condition respectively.

He pointed to the delicate feelings or emotions 'quite tired' and 'very tired' for tiredness in the first and second place respectively. He then pointed location for the delicate feelings 'a little tired' and 'somewhat tired' almost middle of analog scale respectively. He pointed location on the analog scale for 'not tired at all' in the last position. The delicate feelings or emotions 'Not a little tired' is pointed too close as 'not tired at all'. Finally, he pointed the delicate feeling or emotions 'not too tired' and 'not so tired' far from 'somewhat tired' and near the place 'not a little tired'. He thought that 'not too tired' and 'not so tired' are very close to him to convey tiredness.

F. Improvement of Design Solutions

Fig. 9 shows the design of the existing prototype to convey delicate feelings or emotions for physical conditions and tiredness. The existing prototype can be improved by the proposed design as shown in Fig. 10. To improve the prototype, the pointed location is used to support the aphasia participant who wants to convey his delicate feelings or emotions to others. He pointed location on an analog scale to convey his delicate feelings or emotions for physical conditions and tiredness are shown in Fig. 7 and Fig. 8.

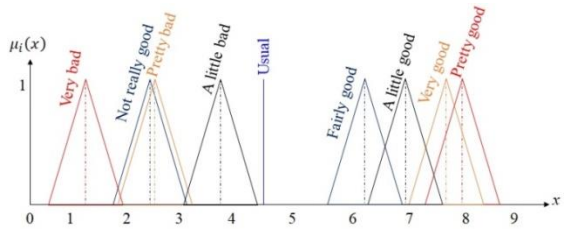


Fig. 7. Delicate Feelings or Emotions in Triangular Fuzzy form to Convey Physical Conditions.

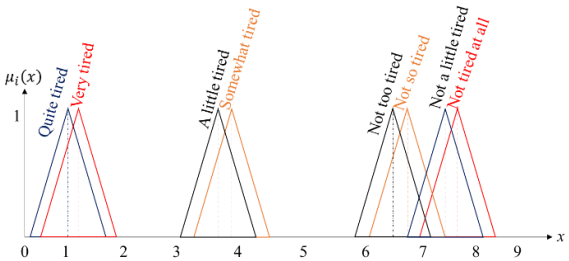


Fig. 8. Delicate Feelings or Emotions in Triangular Fuzzy form to Convey Tiredness.

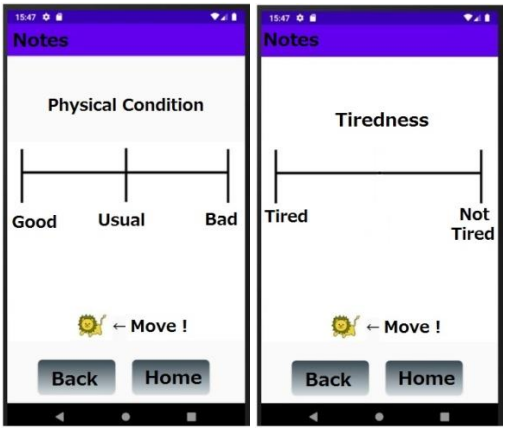


Fig. 9. Existing Design of Physical Condition (Left) and Tiredness (Right) to Convey Delicate Feelings or Emotions.

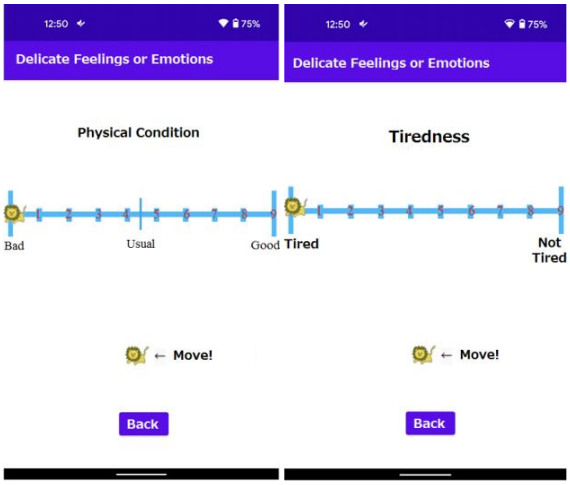


Fig. 10. Proposed Design of Physical Condition (Left) and Tiredness (Right) to Convey Delicate Feelings or Emotions.

The delicate feelings or emotions will be displayed based on the pointed location on the analog scale of the improved prototype for case study as shown in Fig. 11 and Fig. 12. For example, aphasia participant feels 'Pretty bad' and he wants to convey his delicate feelings or emotions to others based on his difficulties. Thus, he points an area between 'bad' and 'good' physical conditions, then the possible candidates of delicate feelings or emotions are selected based on the pointed location as shown in Fig. 11(a). The possible candidates of delicate feelings or emotions and their order are regarded as 'Not really good' and 'Pretty bad' for the pointed area as shown in Fig. 11(a). His desired candidate 'Pretty bad' is placed in the second position on the candidate list. For this reason, he wants to display the list in a more accurate manner. Therefore, aphasia participant moves the pointed location for 'Pretty bad' to $x+\Delta x$ to adjust his delicate feelings or emotions and the order of delicate feelings or emotions are changed to 'Pretty bad' and 'Not really good' as shown in Fig. 11(b). In this way, he gets the desired delicate feelings or emotions at the top of the list.

Furthermore, aphasia participant feels 'Not too tired' and he wants to convey his delicate feelings or emotions to others based on his difficulties. Thus, he points an area between 'tired' and 'not tired' for tiredness, then the possible candidates of delicate feelings or emotions for tiredness is selected based on the pointed location as shown in Fig. 12(a). The possible candidates of delicate feelings or emotions are regarded as 'Not so tired' and 'Not too tired' for the pointed area as shown in Fig. 12(a). He found that the list of possible delicate feelings or emotions do not have his desired item. Therefore, aphasia participant moves the pointed location for 'Not too tired' to $x+\Delta x$ to adjust his desire delicate feelings or emotions, the delicate feelings or emotions are changed to 'Not too tired' and 'Not so tired' respectively as shown in Fig. 12(b). Now, he gets his desired delicate feelings or emotions to convey the tiredness.

Fig. 11. Adjustment of Delicate Feelings or Emotions for Physical Conditions: (a) Conveying Delicate Feelings or Emotion Normally (b) The Delicate Feelings or Emotions after moving the Pointer Δx.

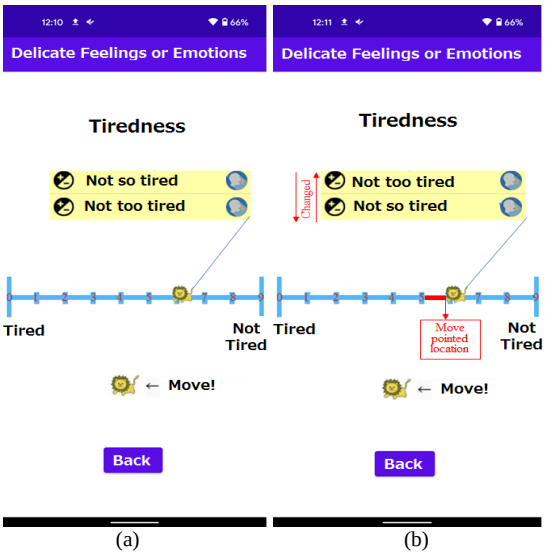


Fig. 12. Adjustment of Delicate Feelings or Emotions for Tiredness: (a) Conveying Delicate Feelings or Emotion Normally (b) The Delicate Feelings or Emotions after Moving the Pointer Δx.

G. Evaluation of the AAC systems

In this study, the evaluation was performed by the first aphasia individual who identified the interaction problems and provided his insights for the redesign. In evaluation process, interfaces of the prototype for physical condition after the redesign and before the redesign were provided to the first aphasia individual. We asked his opinions through a survey questionnaire about the improved interfaces after the redesign compared to before the redesign as shown in Table IV.

For conveying delicate feelings or emotions by the first aphasia individual after the redesign, the responses for the questionnaire were collected on a percentage scale (0 to 100). 100% means that he can completely convey and understand the delicate feelings or emotions to participate in society after the redesign. 0% means that his level of conveying and understanding for delicate feelings or emotions did not improve after the redesign.

The questionnaire was designed in two categories. The first category was for the quantitative opinions of the first aphasia individual to convey delicate feelings or emotions for physical conditions. The second category was for the quantitative opinions for understanding delicate feelings or emotions for physical conditions. In addition, open-ended questions were also attached to each question. After collecting his responses, it is easily determined which interfaces can sufficiently useful to convey delicate feelings or emotions.

As shown in Table IV, the first aphasia individual responded that he could properly convey 50% delicate feelings or emotions for the physical conditions to others with the improved interfaces after the redesign compared to before the redesign. He also believed that he could actively participate in society 30% by conveying delicate feelings or emotions to others with the improved interfaces after the redesign. Moreover, he could properly understand 70% delicate feelings or emotions of others with the improved application after the redesign. Furthermore, he thought that others could understand 70% of his delicate feelings or emotions with the improved interfaces after the redesign.

TABLE IV. SURVEY QUESTIONNAIRE FOR EVALUATION OF THE PROPOSED DESIGN OF AAC SYSTEMS AND RESPONSES OF THE FIRST APHASIA INDIVIDUAL

Questions	**Responses**
How much properly can you convey your physical conditions to others?	I can properly convey 50% delicate feelings or emotions to others
How much more can you actively participate in society by conveying delicate feelings or emotions for physical conditions to others using the improved interfaces?	I can actively participate in society 30% by conveying delicate feelings or emotions to others
How much properly can you understand the physical conditions of others using the improved interfaces?	I can properly understand 70% delicate feelings or emotions of others
How much properly can others understand your physical conditions using the improved interfaces?	Others can understand 70% of my delicate feelings or emotions

IV. DISCUSSION

People with aphasia individuals cannot convey their delicate feelings or emotions with the existing design of AAC systems [15-19]. They used the existing AAC systems only for expressing a few words or phrases which cannot be conveyed by them instantly. On the other hand, the experimental result of this study showed that the first aphasia individual can be expressed more delicate feelings or emotions to others using the proposed design of AAC systems. As shown in Table IV, he can convey 50% delicate feelings or emotions for physical conditions with the proposed design of AAC systems because the design of existing prototypes is improved based on his situations. He suffered from Broca's aphasia with the problem of memory impairment caused by two strokes. He can communicate with others, but his speech is non-fluent because he cannot process the languages properly in his brain. Although he could convey his delicate feelings or emotions for physical conditions in some extent to others during communication, his expressions were not understood properly by others. Due to the memory impairment and the language processing problems of Broca's aphasia, he cannot find sometimes any words/proper words in his communication when he uses existing AAC systems. As a result, he cannot verbally express his delicate feelings or emotions properly to others with the existing AAC systems. Using the proposed design of AAC systems, candidate(s) of delicate feelings or emotions were shown based on his specified position x on the scale. If the candidates were unable to properly express his delicate feelings or emotions, then the function of conveying the delicate feelings and emotions can be realized by controlling the position of the fuzzy membership function from x to $x+\Delta x$ or $x-\Delta x$. The position of delicate feelings or emotions was difficult to control without the help of Fuzzy-set theory. As a result, the participant can express his physical conditions to others more properly with his delicate feelings or emotions.

The first aphasia individual cannot properly express his own thinking to others because he was Broca's non-fluent aphasia. He had no problems with hearing and understanding, but it was difficult to him to speak the contents of communication. The difficulty of speaking becomes the barrier of his communication before the proposed design. As shown in Table III, he could convey delicate feelings or emotions for physical conditions in some extent with his barrier. Thus, interfaces design for physical conditions using smartphone applications before and after the proposed design have been evaluated. As shown in Table IV, the experimental result showed that the proposed AAC systems design is improved because the participant can convey properly 50% delicate feelings or emotions to others through the proposed AAC systems. This improvement was found after the proposed design because the possible candidate(s) of delicate feelings or emotions were shown using Fuzzy-set theory based on the priority. The order of candidate(s) of delicate feelings or emotions was changed after managing Δx on the scale. Δx was managed so that fuzzy membership function fits more based on his situation. The management of Δx was very helpful to him because it became a supporter for him. Consequently, he can convey 50% delicate feelings or emotions for physical conditions with the proposed design of AAC systems as shown

in Table IV. Therefore, Fuzzy-set theory in design of AAC systems contributed to properly convey the delicate feelings or emotions to others.

V. CONCLUSION

This study presents the design of an AAC system to convey delicate feelings or emotions for aphasia individuals by the support of Fuzzy-set theory. The design was started from the interaction problems between aphasia individuals and existing AAC systems. To identify the interaction problems, a survey questionnaire is provided to the aphasia individuals in three categories such as conveying delicate feelings or emotions, conveying physical conditions, and conveying tiredness. In addition, an existing prototype of AAC systems is also provided to the aphasia individuals. Moreover, an aphasia individual specified fuzzy membership functions for delicate feelings or emotions as his insights during the redesign. Based on the Fuzzy-set theory, the prototype was improved to convey his exact feelings or emotions to others in daily life. Using the proposed design of AAC systems, conveying delicate feelings or emotions of the first aphasia individual for physical conditions to others can be improved by 50%. As a result, other disabled individuals like him can also convey their delicate feelings or emotions using the proposed design of AAC systems. However, this study investigates how the AAC systems is designed for conveying physical conditions and tiredness with the delicate feelings or emotions to others using Fuzzy-set theory. In the future, other factors will be investigated to apply Fuzzy-set theory at design of AAC systems for conveying delicate feelings or emotions.

REFERENCES

[1] Doogan, C., Dignam, J., Copland, D. and Leff, A., "Aphasia recovery: when, how and who to treat?," Current neurology and neuroscience reports, vol. 18, no. 12, p.90, 2018.

[2] Pedersen, P.M., Stig Jørgensen, H., Nakayama, H., Raaschou, H.O. and Olsen, T.S., "Aphasia in acute stroke: incidence, determinants, and recovery," Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society, vol. 38, no. 4, pp.659-666, 1995.

[3] Laska, A. C., Hellblom, A., Murray, V., Kahan, T. and Von Arbin, M., "Aphasia in acute stroke and relation to outcome," Journal of internal medicine, vol. 249, no. 5, pp. 413-422, 2001.

[4] Beukelman, D. R. and Mirenda, P., "Augmentative & alternative communication: supporting children & adults with complex communication seeds," Paul H. Brookes Publishing Company, 2013.

[5] Vance, A., "Insurers fight speech-impairment remedy," The New York Times. September 14, 2009.

[6] Assistiveware B.V Proloquo2Go (2017), Proloquo2Go official website http://www.assistiveware.com/product/proloquo2go, Accessed January 30, 2022.

[7] ANSI A117.1-1961, American Standard Specifications for Making Buildings and Facilities Accessible to, and Usable by, the Physically Handicapped, 1961.

[8] Berube, B., "Barrier-free design-making the environment accessible to the disabled," Canadian Medical Association Journal, vol. 124, no. 1, p.68, 1981.

[9] Mace, R., "Universal design: Barrier free environments for everyone," Designers West, vol. 33, no. 1, pp.147-152, 1985.

[10] EIDD: The EIDD Stockholm Declaration 2004, at the Annual General Meeting of the European Institute for Design and Disability in Stockholm. Design for All Europe, 2004.

[11] Harper, S., "Is there design-for-all?," Universal Access in the Information Society, vol. 6, no. 1, pp.111-113, 2007.

[12] Norman, D. A. and Draper, S. W. (Editors), "User-centered system design: new perspectives on human-computer interaction," Lawrence Earlbaum Associates, Hillsdale, NJ, 1986.

[13] Karlsson, J., & Ryan, K., "A Cost-Value approach for prioritizing requirements," IEEE Software, vol. 14, no. 5, pp. 67-74, 1997.

[14] Norman, D., "The design of everyday things (Revised and expanded edition)," USA, Basic Books, pp. 219, 2013.

[15] Al Mahmud, A. and Martens, J.B., "Amail: design and evaluation of an accessible email tool for persons with aphasia," Interacting with Computers, vol. 25, no. 5, pp.351-374, 2013.

[16] Allen, M., McGrenere, J. and Purves, B., "The field evaluation of a mobile digital image communication application designed for people with aphasia," ACM Transactions on Accessible Computing (TACCESS), vol. 1 , no. 1, pp.1 -26, 2008.

[17] Koppenol, T., Al Mahmud, A., and Martens, J. B., "When words fall short: helping people with aphasia to express," In International Conference on Computers for Handicapped Persons, Berlin, Heidelberg, pp. 45–48, 2010.

[18] Kane, S.K., Linam-Church, B., Althoff, K. and McCall, D., "What we talk about: designing a context-aware communication tool for people with aphasia," In Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, ASSETS '12, New York, NY, USA, pp. 49–56, 2012.

[19] Hossain, M.S., Takanokura, M., Sakai, H. and Katagiri, H.,. "Using context history and location in context-aware AAC systems for speech-language impairments," In Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, 2018.

Detecting Ransomware within Real Healthcare Medical Records Adopting Internet of Medical Things using Machine and Deep Learning Techniques

Randa ELGawish[1]
Mohamed Hashim[4]
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Mohamed Abo-Rizka[2]
College of Computing and Information Technology
Arab Academy for Science and Technology, Cairo, Egypt

Rania ELGohary[3]
Artificial Intelligence Technology Center, Misr University for Science and Technology,
Cairo, Egypt

***Abstract*—The Internet of Medical Things was immensely implemented in healthcare systems during the covid 19 pandemic to enhance the patient's circumstances remotely in critical care units while keeping the medical staff safe from being infected. However, Healthcare systems were severely affected by ransomware attacks that may override data or lock systems from caregivers' access. In this work, after obtaining the required approval, we have got a real medical dataset from actual critical care units. For the sake of research, a portion of data was used, transformed, and manifested using laboratory-made payload ransomware and successfully labeled. The detection mechanism adopted supervised machine learning techniques of K Nearest Neighbor, Support Vector Machine, Decision Trees, Random Forest, and Logistic Regression in contrast with deep learning technique of Artificial Neural Network. The methods of KNN, SVM, and DT successfully detected ransomware's signature with an accuracy of 100%. However, ANN detected the signature with an accuracy of 99.9%. The results of this work were validated using precision, recall, and f1 score metrics.**

Keywords*—Artificial neural networks; deep learning; healthcare system; internet of things; machine learning; supervised learning

I. Introduction

The Internet of Medical Things (IoMT) is a collection of medical devices and applications that use networking technologies to connect to clinical information systems. It can reduce unnecessary hospital visits and the burden on healthcare systems by connecting patients to their medical practitioners and allowing their medical data to get transferred over a secured network. According to Frost & Sullivan, the global IoMT market was worth $22.5 billion in 2016 and was expected to be worth $72.02 billion by 2021, at a compound annual growth rate of 26.2 % [1].

According to NBC News, ransomware malware severely infected a major hospital chain in September 2020. Its impacts caused all employees and medical staff to be forced to use the traditional pen and paper method to monitor the patient's status over the weekend. This cyber-attack became the most significant in the history of the United States, as it has affected over 400 locations [2].

Ransomware is a form of malware designed to encrypt data partially or as a whole, causing the systems that rely on them to become unusable. Ransomware exists in two types; crypto and locker. According to Kaspersky [3], crypto-ransomware encrypts valuable files on a computer, making them inaccessible to the user.

Cybercriminals who carry out crypto-ransomware generate profit by demanding victims pay a ransom to recover their files. However, paying the ransom never guarantees the recovering of the victim's files. In crypto-ransomware, files are not encrypted by locker ransomware, but instead, it locks the victim out of their device(s), making it inoperable. Once locked out, cybercriminals will start executing inside attacks pressuring the victim to pay a ransom to unlock their device(s).

Crypto-ransomware can get subcategorized into other types, including payload ransomware which is the research focus in this paper. Payload ransomware encrypts values randomly stored within valuable files on a computer. In healthcare systems, encrypting values, especially for intensive care units (ICU), could lead to the loss of lives. IoMT has served the caregivers and healthcare during the covid 19 pandemic heavily as it has minimized the contact between the hospital staff and their patients. However, they need an indeed security against such attacks. In this paper, our primary focus is to detect payload ransomware's signature that has infected our medical data named Mimic III v1.4 [4]. Section IV will reveal details about this medical data and how the appropriate access is granted to researchers. Machine and deep learning techniques became viral tools that have attracted the attention of researchers in data analytics and cyber-security domains. However, the data collected from the perception layer is accompanied by numerous complications, such as dynamic data changes, their large volume accompanied by noises. These challenges require developing and implementing efficient methods to validate, visualize and extract knowledge from this immense amount of data.

This research paper is organized as follows. A literature review of other researchers detecting cyber-attacks in IoT is presented in Section II, followed by the methodologies used to detect the encrypted medical records within the dataset in Section III. Section IV briefly explained the medical dataset

used and prepared for the detection by the supervised learning techniques after feature selection, manifestation, clustering, and preprocessing. The detection of ransomware-infected records by machine learning and deep learning is explained in Section V. The evaluation of results and discussion are described in Section VI. Lastly is the conclusion in Section VII.

II. Related Work

M. M. Rashid, J. Kamruzzaman, M. Hassan, T. Imam, and S. Gordon, in [5], implemented decision trees, random forest, linear regression, support vector machine, and artificial neural network to detect cyber-attacks at fog nodes within a distributed rather than centralized system to track network traffic using two different datasets UNSW-BC15 and CICI052017. Their experimental results showed DT and RF performed better in terms of accuracy than the other algorithms.

In another study [6], M. Hasan, M. Islam, I. Zarif and M.M.A Hashem used a publicly accessible IoT dataset [7] and proposed a data analysis method to identify and prevent systems from attacks that cause abnormal behavior. They used DT, RF LR, SVM, and ANN; however, RF scored the best accuracy.

In [8], A. A. Diro and N. Chilamkurti proposed a deep learning model versus a shallow neural network model to detect normal, DoS, probe R2L and U2R traffic using the NSL-KDD dataset. The two models scored the following accuracies respectively, 99.2 % versus 98.27 % for binary classification and 95.22 % versus 96.75 % for multi-class classification.

R. Doshi, N. Apthorpe and N. Feamster in [9] trained binary classifiers to differentiate between benign and denial of service (DoS) attack traffic generated by mirai botnets. The results showed a good performance in detecting well-known signature attacks compared to new or unknown ones.

In [10], B. Ingre and A. Yadav applied ANN to NSL-KDD dataset to analyze binary and five class classification. The detection accuracy was 81%, in addition to 79 % in detecting the attack classification type.

Moreover, a hybrid learning model proposed by M. M. Lisehroodi, Z. Muda and W. Yassin in [11], implemented ANN and K-means methods for clustering within their design for an advanced network intrusion detection system. Their results showed a successful detection accuracy of 99 %.

Another hybrid model [12] was proposed by S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, in which DT and SVM are implemented and compared to each one of them individually. Their experimental results showed DT has equal or slightly better performance in comparison to SVM and DT-SVM.

J. Zhang and M. Zulkernine in [13] implemented RF in network intrusion detection systems against DoS attacks to overcome imbalanced intrusions and reduce the error rate from 1.92 % to 0.05 %.

In this study [14], S. Mukkamala, G. Janoski and A. H. Sung applied a strategy that uses ANN and SVM to detect network traffic. ANN had an accuracy of detection of 99 %; however, the training time was 30 minutes and further 30 minutes for testing. While SVM had a lower accuracy score, it took around 52s to 211s for training and other 1s to 16s for testing.

In addition to another research where A. Azmoodeh, A. Dehghantanha, M. Conti and K.Choo target the detection of crypto-ransomware via these learning methodologies and monitor the power consumption of internet-connected devices or android devices [15].

Our work measures the accuracy of detecting payload ransomware infected patients' records within our obtained medical dataset using supervised machine learning and deep learning techniques.

However, the availability of such medical datasets is highly restricted and inaccessible unless approved by the appropriate parties after fulfilling the Health Insurance Portability and Accountability Act (HIPAA) standards and signing a data user agreement (DUA).

Therefore, studying some of the impacts of ransomware on medical data is highly demanded, significantly when this malware negatively impacts them.

III. Methodology Framework

In the following Fig. 1 is an illustrated overview of our suggested IoMT network. In this network, devices are portable and dispersed within a network-defined range, with the edge router serving as a coordinator between the control and IoMT environment zones. Wireless communication protocols are used by devices to communicate to the server in the control zone.

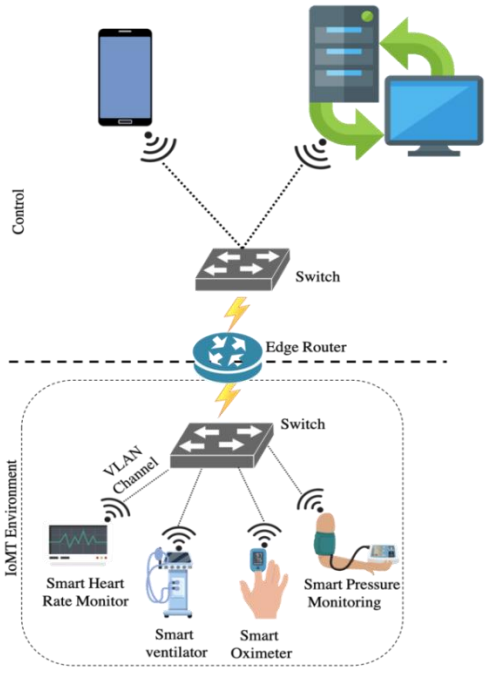


Fig. 1. System Overview.

A. Proposed Detection System

This section will discuss the steps taken to get the dataset prepared for being labeled after manifestation. This step is followed by the detection phase using the previously mentioned machine and deep learning techniques. The proposed mechanism works as illustrated in Fig. 2. Our dataset is composed of more than 300 different physiological tests available to be performed on patients during their ICU stay period.

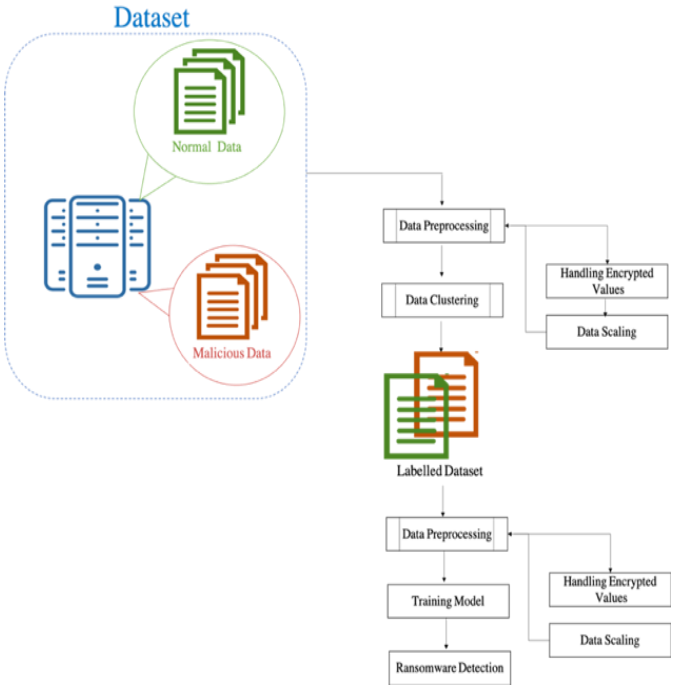


Fig. 2. Proposed System for Detection.

Therefore, we have chosen the most frequent tests performed in the ICU as our features. As the rest of the tests are not in demand or necessary to be conducted, they are performed depending on their patients' medical condition.

These tests are eligible to be captured via sensors and bedside monitors to orchestrate an IoMT environment. After transformation, the dataset becomes manifested using payload ransomware that is described in the following section. This step involves selecting random records and encrypts some of their respective features at random.

Any record that has any of its fields encrypted is considered infected. The records are either labeled as infected by number 1 or 0 for the normal ones in an additional column. This label column is later used to evaluate the supervised machine learning and deep learning techniques implemented in this work.

IV. Dataset Preparation

A. Original Dataset

The clinical dataset, MIMIC-III v1.4, used in this work is a credential actual medical data obtained from Beth Israel Deaconess Medical Center, Boston, MA, United States. This data was accessed after becoming a credentialed user on PhysioNet.

This step involves the completion of a training course on human subject research. The training received were "Human Resource: Data or Specimens Only Research" and "Human Research Data or Specimens Only Research (Course Learner Group 2).

These training courses ensure that the researchers seeking the use of this database would treat it with care and respect since it contains detailed clinical care information about patients, in addition to signing a DUA.

MIT Laboratory for Computational Physiology removed patients' critical health information such as diagnostic reports or text fields from the database. They fulfilled the HIPAA standards by eliminating all data elements such as the patient's name, phone number, and address. A random offset was used to shift the dates into the future. In addition to hiding the patients' actual age by moving it, some of this data showed the patients over 300.

The dataset was obtained using two different clinical information systems CareVue and MetaVision. Some of the tests were recorded but under other names according to their associated clinical information system.

After inspecting the dataset, it was found that some tests were monitored by one of the clinical information systems while the other did not. Therefore, it was necessary to manually go through the dataset to choose the features, especially those under shorthand terms.

The dataset consists of 61,532 ICU stays, of which there are 53,432 stays for adult patients and 8,100 for neonatal patients; however, this doesn't affect the features chosen for this work. The whole dataset is composed of 29 distinct tables.

The table of interest stores all of the medical tests conducted during the patient's ICU stay. As previously mentioned, the medical tests were under shorthand terms. Therefore, using a table provided in the dataset that provides the complete form of these tests aided in identifying the concept measured.

The choice of the features depends on its frequency; the number of the concept was carried out by devices. In addition to its eligibility to be captured remotely without any external factor, i.e. caregiver, to promote an actual remote monitoring environment. The number of records in this table was over 310 million records. Therefore, as a proof of concept, one million records were used for this investigation.

B. Feature Selection

Computing the frequency of each medical concept measured in our table of interest has resulted 353 different medical tests. Their frequencies range from 1 to 8263. Therefore, the chosen medical features are the features that have number of occurrences above 5000.

In contrast, the rest of the features had frequencies ranging from 1 to 2000. In Table I, the chosen six features are presented.

TABLE I. LIST OF FEATURES SELECTED

Feature n	*Feature Name*
1	Heart Rate
2	Respiratory Rate
3	O_2 Saturation Pulseoxymetry
4	Non-invasive blood pressure systolic
5	Non-invasive blood pressure diastolic
6	Non-invasive blood pressure means

The heart rate had a frequency of 8263. The respiratory rate was 8213. 8148 was for o_2 saturation pulseoxymetry was 8148. The non-invasive blood pressure systolic, diastolic and means are; 5526, 5524 and 5559. In this dataset, a patient could have had any of these features numerous times a day, e.g. 30 and extremely few had the first three features measured only. For those who had certain features tested innumerable times, the means of these values were computed for every single day instead.

Note that each record is associated with additional fields such as date, time and ICU ID. The ICU ID column stores the unique numbers given to the patients once admitted into the ICU. The feature's string name was used, such as heart rate in the record itself, to represent that it has been captured along with its value in the CSV file. Therefore, we have rearranged the table into transpose, and all of these features became the headers of the CSV file. This transformation has led to a decrease in the number of records within the file to 149, 360 records.

C. Data Manifestation and Labeling

The data was split into two portions (51 % and 49%). The more significant portion was fed into the manifestation process. Some of the fields' values in the patients' records were encrypted randomly regardless of their data type using the algorithm shown in Fig. 3, hiding their valid values under the signature "□ PayMeLocker Decrypt □".

Note that the malware did not make the whole record encrypted; some of its fields were encrypted, including the date, time and ICU ID. A record is considered infected if one or more fields are encrypted. Thus, this step was followed by clustering to label the record as 1 or 0 depending on its manifestation case.

Both manifested, and regular records were merged again at random; ready to be labeled. A threshold-based method was used to label the records.

We have handled the encrypted values by replacing them with unique constant numbers depending on the feature infected within the respective record. The constant value for each feature is shown in Table II.

Algorithm 1: Payload Ransomware Manifestation

Require: Data in Transpose;

1: **function** INFECT (file)
2: *Data* ← *file*
3: **for each** *r* ∈ *Data* **do**
4: **for each** *f* ∈ *Data* **do**
5: Generate *state* /* *0 or 1 state*
6: **if** state = *0* **then**
7: *f*++
8: **else**
9: Data [*r*][*f*] = Encrypt the value stored in the *feature*
10: **end if**
11: **end for**
12: **end for**
13: **end function**

Fig. 3. Manifestation Algorithm.

Note that the specific number for each feature was decided by finding the maximum value found in the dataset for each feature and doubling it to ensure its differentiation compared to the rest of the values within the same record. The first three features are in Table II represent the date, time, and ICU ID.

This step is followed by data scaling. Data scaling is a technique that normalizes the data values of features within a given dataset into a particular range. After replacing the encrypted values with numerical values, the data within each column were normalized to floating numbers ranging from zero to 1.

TABLE II. LIST OF THE DISTINCT VALUES WITHIN EACH FEATURE

Feature n	*Maximum Value*	*Distinct Value*
1	26758	53516
2	23.5	47
3	299707	599414
4	200	400
5	265	530
6	193	386

This step was implemented using the library of preprocessing and MinMaxScaler (). If the value under the selected feature carries a numerical value equivalent to 1, the whole record becomes labeled as one, i.e. infected; otherwise, it is labeled as zero, i.e. normal.

V. DETECTION

This section will discuss the procedure used for training and detecting the infected records using the most commonly implemented supervised machine and deep learning methods in the internet of things.

A. Detection using Machine Learning Techniques

To implement KNN, SVM, DT, RF, and LR, we have used their python-based libraries of sklearn neighbors, SVM, tree, ensemble, and linear model. The training dataset was read, stored in a data frame, and converted into a matrix during the classification stage.

Furthermore, these datasets are divided into training and testing datasets. The training dataset being the larger is composed of 45, 034 bengin records and 47,466 infected ones. In comparison, the testing dataset is composed of 14369 bengin records and 15503 infected ones.

After the initial training of the machine learning models using the labeled training dataset, it was applied to the testing dataset with no binary label information. Note that the k-fold cross-validation procedure was applied as a part of this work using train/test split to avoid overfitting.

B. Detection using Deep Learning Techniques

The deep–learning model of ANN uses supervised training and binary classification for identifying the infected tuples. A four-layer deep learning model was created for this study, with one input layer, two hidden ANN layers, and one output layer, a binary classifier layer.

The input layer consists of nine neurons while the hidden layers; each consist of eight neurons, and lastly, the output layer comprises two neurons. Each neuron in the ANN layer is assigned with a weight parameter adjusted using the gradient descent method, Fig. 4.

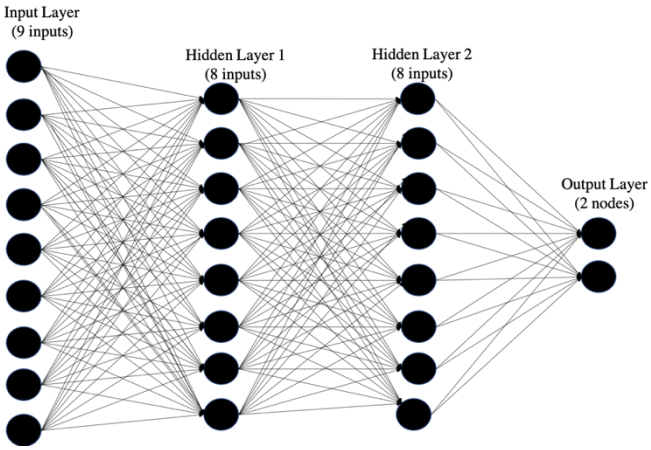


Fig. 4. Detection Mechanism using Neural Network.

Each tuple and its label information are fed into the ANN during the supervised training process, passing through the first hidden encode layer and being filtered out as x, the most significant features. The x features are passed into the second encode hidden layer, filtered and converted into y features.

Finally, the second encode layer sends them to the output layer, where they are classified as malicious or benign tuples. This process takes place at the lowest level when the feature vector fv is input into the ANN, and it passes through each layer of the deep neural network (DNN), Fig. 5.

Each DNN layer's neural nodes calculate an output using an activation function and generate a filtered result. In this work, we developed this model using a rectified linear unit (ReLU) activation function. The ReLU function is defined as follows:

$$f(x) = \max(0, x) \quad (1)$$

The smaller values in the matrix are set to zero with the input x (i.e. matrix), while the others remain constant. As a result, each hidden layer connects to the next hidden layer via a linear combination of outputs and feeds the filtered output generated by the ReLU activation function to the next layer.

The second encoded layer, like the first encoded layer, trains itself using labeled tuples. As a result, each layer of the ANN feeds on this data and maps it to a numerical value. Finally, the mapped values are normalized to 0 and 1, with 0 representing the benign tuple and 1 representing the malicious tuple.

The ANN model's objective function, a binary_crossentropy loss function, tries to minimize the total cost in the model, as shown in the following algorithm, Fig. 6. The ANN model must then be retrofitted for training and testing predictions.

The ANN was implemented using Keras, an open-source neural network library written in python, and the results are validated as well using the confusion matrix. During the classification stage, the training dataset was read, stored in a data frame in the same way as in the machine learning model. Furthermore, the same training dataset and testing dataset were used in this implementation to deduce the accuracy performance of this model with the previously conducted machine learning algorithms.

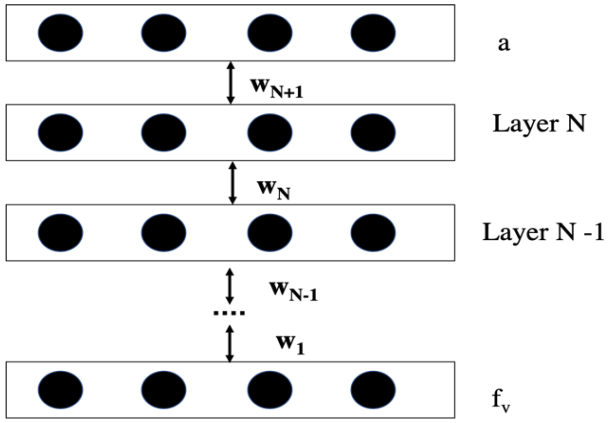


Fig. 5. DNN Structure.

Algorithm 2: Ransomware Detection using Deep-Learning model

Require: List of the 9 features

1: **function** DETECT (file)
2: *matrix* ← *file*
3: Extract *features* from matrix
4: Define $dataset_{train}$ & $dataset_{test}$
5: Initialize *sequential deep-learning model*
6: **if** *initialized* **then**
7: Compile *binary-crossentropy* classifier
8: classifier ← *sequential deep-learning model*
9: **end if**
10: Training: classifier: $dataset_{train}$
11: **if** *Training* is complete **then**
12: Prediction: classifier ← $dataset_{test}$
13: **if** *Predictions* are correct **then**
14: Re-Train the *model*
15: **end if**
16: **end if**
17: **end function**

Fig. 6. Detection Mechanism using Deep Neural Network.

A sequential 2 hidden layer was created and instantiated the ANN model, and the ReLU activation function was used to equip the processing units within each layer. Following that, the deep learning model was compiled and fitted with 100 runs, i.e. epochs and feature count. Finally, the deep learning model is assembled, and the classifiers are assigned and saved in the variable "prediction". Consequently, in every test, the results of the classifier are normalized into a binary value.

VI. Evaluation and Results

The performance metrics used to evaluate the detection model are; recall, precision and f1 score. Precision(P) and recall (R) are two essential metrics used to assess the accuracy of the detection process when there is an imbalanced classification.

These two metrics use true positive value as an outcome when the model predicts the positive class correctly. A true negative, on the other hand, is an outcome in which the model correctly predicts the negative class. A *false positive*, on the other hand, is an outcome in which the model mispredicts the positive class. A false negative is an outcome in which the model mispredicts the negative class. Lastly, a presented P-R curve refers to the composition of these two metrics.

The precision is referred to as the positive predictive value and outlines how good a model predicts the positive (anomaly) label. To calculate precision, we use the following formula:

$$P=TP/(TP+FP) \quad (2)$$

The recall is the ratio between the number of true positive labels divided by the sum of the true positive values and the false negative values. To calculate recall, we use the following formula:

$$R=TP/(TP+F) \quad (3)$$

The f1 score is a measure of a test's accuracy. It depends on the values of precision and recall. To calculate the F1 score, we use the formula below:

$$F1=(2(P)(R)) /((P+R)) \quad (4)$$

Table III shows the number of TP values in KNN, 15,453, while the true negative values are 14,419 with zero errors. These numbers mark the true actual percentage of the benign and malicious tuples within the dataset.

KNN has shown excellent performance in the detection processes with 100 % precision and recall, as shown in Table IV; however, these percentages are expected to decrease with the increase in dimensionality. KNN can aid in the detection process if accompanied by a principle component analysis (PCA) algorithm.

SVM has reached an overall precision and recall of 96 % compared with the rest of the methods in this work. Thus, SVM can show excellence when the training dataset is not large, which is not the case in biomedical data.

DT and RF did reach 100 % in terms of accuracy; however, it was computationally expensive in the word of memory space. LR has the lowest precision and recall rates, marking the worst percentage compared to the other algorithms.

ANN has scored 99.9 % in precision and recall and is expected to reach 100% as the dataset dimensionality increases. This score proves that ANN can be used and furtherly developed to be able to detect ransomware signatures.

TABLE III. Table of TP, TN, FP and FN Values for the Detection Techniques Implemented

Technique	***TP***	***TN***	***FP***	***FN***
KNN	15453	14419	0	0
SVM	14558	14012	563	739
DT	15304	14568	0	0
RF	15331	14541	0	0
LR	14693	13322	1273	584
ANN	15305	14545	11	0

TABLE IV. TABLE OF PRECISION, RECALL AND F1 SCORE VALUES FOR THE DETECTION TECHNIQUES IMPLEMENTED

Technique	*Precision*	*Recall*	*F1*
KNN	100 %	100 %	100 %
SVM	96.0 %	96 %	96 %
DT	100 %	100 %	100 %
RF	100 %	100 %	100 %
LR	92.0	96.7	94.0%
ANN	99.9%	100 %	99.5%

The following P-R curve compares the detection performances of ransomware detection using the same dataset and same training and testing percentages, Fig. 7.

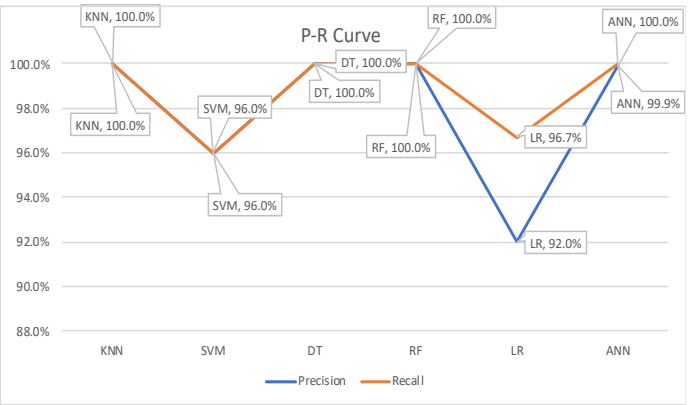


Fig. 7. P-R Curve.

In this work, we have detected infected ransomware tuples using supervised machine and deep learning techniques of KNN, SVM, DT, RF, LR, and ANN. The encryption signature of the malware used is evident within the dataset and very different from the original values stored under their respective features. Therefore, the precision score of KNN, DT, RF was 100 %, and ANN had it almost there.

VII. CONCLUSION

In this paper, machine and deep learning techniques were used to perform binary classification on a medical dataset infested with payload ransomware. The Healthcare system was not on the top priority for the security specialists a few years ago. Until the emergence of the Internet of Medical things that was heavily implemented during the pandemic of Covid 19 to enhance the infection control process. In addition to the immense increase of sensitivity of the data, it was transferring. When it was exposed to catastrophic attacks, especially ransomware, it was time to get experiment with the effectiveness of using these learning methods that have been used to famous attacks like DDoS and DoS, in detecting payload ransomware on real healthcare datasets. A subset of the dataset used was cleaned, transformed, and infested with the attack. The work implemented was evaluated by the recall, precision, and f1 score metrics. The results showed ANN showed 99.9% of accuracy in detection even while KNN, SVM, and DT were 100%. These results show that these methods can be considered to secure the data within medical or healthcare systems.

REFERENCES

[1] Internet of Medical Things Revolutionizing healthcare, Alliance of Advanced BioMedical Engineering. Available Online: https://aabme.asme.org/posts/internet-of-medical-things-revolutionizing-healthcare (accessed on 2017).

[2] K. Collier, NBC News. Available Online: https://www.nbcnews.com/tech/security/cyberattack-hits-major-u-s-hospital-system-n1241254 (accessed on September 2020).

[3] Ransomware Attacks and Types – How Encryption Trojans Differ, Kaspersky. Available Online: https://me-en.kaspersky.com/resource-center/threats/ransomware-examples (accessed on 2021).

[4] A. Johnson et al., MIMIC-III, a freely accessible critical care database. Scientific Data 2016**,** vol. 3, DOI: 10.1038/sdata.2016.35.

[5] M. M. Rashid, J. Kamruzzaman, M. Hassan, T. Imam and S. Gordon, Cyberattacks Detection in IoT-Based Smart City Applications Using Machine Learning Techniques. Int. J. Environ. Res. Public Health 2020, vol. 17, p. 9347.

[6] M. Hasan, M. Islam, I. Zarif and M.M.A Hashem, Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet Things 2019, vol. 7, p100059.

[7] C. C. Aggarwal, J. Han, J. Wang and P. Yu, A framework for clustering evolving data streams. In Proceeding VLDB Conference, Berlin, Germany, 2003; pp. 81–92.

[8] A. A. Diro and N. Chilamkurti, Distributed attack detection scheme using deep learning approach for Internet of Things. Future Gener. Comput. Syst.,2018,vol. 82, pp. 761–768.

[9] R. Doshi, N. Apthorpe and N. Feamster, Machine Learning DDoS Detection for Consumer Internet of Things Devices, IEEE Security and Privacy Workshops (SPW), San Francisco, CA, May 2018, pp. 29–35, DOI: 10.1109/SPW.2018.00013.

[10] B. Ingre and A. Yadav, Performance analysis of NSL- KDD dataset using ANN. In 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, India, Jan 2015, pp. 92-96, DOI: 10.1109/SPACES.2015.7058223.

[11] M. M. Lisehroodi, Z. Muda and W. Yassin, A hybrid framework based on neural network MLP and K-means Clustering for Intrusion Detection System. In proceedings of the 4th International Conference on Computing and Informatics, Sarawak, Malaysia, 2013, pp. 305 – 311.

[12] S. Peddabachigari, A. Abraham, C. Grosan and J. Thomas, Modeling intrusion detection system using hybrid intelligent systems. Journal of network and computer applications 2007,vol. 30, pp. 114-132.

[13] J. Zhang and M. Zulkernine, Network Intrusion Detection using Random Forests, PST, 2005.

[14] S. Mukkamala , G. Janoski and A. H. Sung, Intrusion detection: support vector machines and neural networks. In proceedings of the IEEE - IJCNN, Honolulu, HI, USA, 2002, pp.1702–1707, DOI:0.1109/IJCNN.2002.1007774.

[15] A. Azmoodeh, A. Dehghantanha, M. Conti and K.Choo, Detecting crypto-ransomware in IoT networks based on energy consumption footprint. Journal of Ambient Intelligence and Humanized Computing, 2018, vol. 9, pp. 1141-1152, DOI: https://doi.org/10.1007/s12652-017-0558-5.

Data Visualization of Influent and Effluent Parameters of UASB-based Wastewater Treatment Plant in Uttar Pradesh

Parul Yadav[1], Aditya Chaudhary[2], Anand Keshari[3], Nitish Kumar Chaudhary[4]
Priyanshu Sharma[5], Kumar Saurabh[6], Brijesh Singh Yadav[7]
Department of CSE, Institute of Engineering and Technology, Lucknow, India[1]
Department of CSE, Institute of Engineering and Technology, Lucknow, India[2, 3, 4, 5]
Namami Gange STP Project, Voltas Ltd. Patna, India[6]
Uttar Pradesh Rajya Vidyut Utpadan Nigam Limited, Lucknow[7]

***Abstract*—A rise in the population of a region implies an increase in water consumption and such a continuous increase in the usage of water worsens wastewater generation by the region. This escalation in wastewater (influent) requires the Wastewater Treatment Plants (WWTPs) to operate efficiently in order to process the demand for sewage disposal (effluent). This research paper is based upon visualizing and analyzing the parameters of influent like COD, BOD, TSS, pH, MPN and also, the parameters of effluent like COD, BOD, DO, pH and MPN of Bharwara WWTP situated in Lucknow, India which is the largest UASB-based wastewater treatment plant in Asia. We also design and implement an initial model using the machine learning based techniques to analyze as well as predict the parameters of influent and effluent of the WWTP. Model Performance is measured using Mean Squared Error (MSE) and Correlation Coefficient (R). For analyzing and designing the model, the parameters of influent and effluent have been collected over a period of 26 months on a daily basis covering the variations between seasons and climate. As a result, the model shall provide a better quality of effluent along with consuming the plant resources in an efficient manner.**

Keywords*—Wastewater treatment plant; Bharwara STP; UASB-based plant; influent or effluent prediction; data visualization of influent and effluent; machine learning based for WWTPs

I. Introduction

Wastewater Treatment Plants (WWTPs) take part in playing a crucial role in shaping the urban and rural environments as they are used for processing sewage water and removal of various particles and chemicals which are harmful for the water hydrosphere and the organisms which are dependent on it. An increase in the population of a region implies an increase in water consumption and such a continuous increase in the usage of water results in an increase in the wastewater generated by the region [1]. This increase in influent requires the wastewater treatment plants to operate efficiently in order to process the demand for effluent (sewage disposal) [2, 3, 5].

Besides increase in influent, another more challenging issue in a wastewater treatment plant is the fluctuating or uncertain behaviour of various parameters of the influent in the plant which can be due to varying environmental factors also [15]. To maintain the effluent parameters within the standard range, the wastewater treatment plants need to operate and process on the influent coping up with its varying parameters. On the other side, the wastewater treatment plants require to do optimum utilization of resources during the treatment of influent. Consequently, this uncertain nature of influent parameters demands to find insights and hidden patterns by applying visualization and analytics on the real time historical/ recorded data which in turn shall help to provide/estimate better and efficient (optimized) utilization of resources at wastewater treatment plants. Further knowing the flow and parameters of influent and parameters of effluent in advance shall reduce operational cost of the wastewater treatment plants.

With this objective, we collected and recorded the water parameters for over 26 months (April 2019 to May 2021) from Bharwara wastewater treatment plant situated in Lucknow district which is the largest UASB-based wastewater treatment plant in Asia as it has the capacity to operate and process an average flow rate of 345 Millions of Liter per Day (MLD) with the ability to handle a peak load of 517 MLD of sewage daily. In this paper, we analyze the influent and effluent parameters of Bharwara wastewater treatment plant.

The paper is organized in six sections. Section II presents the research works done in the line of analyzing and predicting influent and effluent parameters across the globe. Section III describes the working of the Bharwara wastewater treatment plant and also its current technological status. In Section IV, we elaborate the methodology of the proposed model. Sections V highlights the results obtained and analysis done. Section VI summarizes the conclusions and shows line of the future works.

II. Related Work

We have carried out the literature survey in the line of research work under two dimensions. The first dimension of study is in the line of exploring technological status of wastewater treatment plants outside India, and the second dimension of study is in the line of exploring technological status of wastewater treatment plants inside India. We shall

discuss both dimensions one by one in the next two subsections.

A. International Status

1) Konya wastewater treatment plant [konya, turkey]: Tümer Abdullah et al. [2], proposed a model using Artificial Neural Network (ANN) for the prediction of Total Suspended Solids (TSS) based on the input parameters Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), TSS. Model performance was evaluated via Mean Squared Error and Correlation Coefficient (R) for the Konya Wastewater treatment plant. Neural Networks of various hidden layers were used and the correlation coefficient in the training set has reached up to 0.99, that is, a satisfactory result for the proposed model. The model was implemented using MATLAB which increased the complexity of designing ANN models as compared to Python. Training and testing take a lot of effort and scaling the models to other WWTP will require ample resources. In this model, the number of layers and number of neurons per layer were decided/ obtained on the basis of analyzing and comparing error on various trails. In the research paper, the performances of nine different models were compared.

2) Wastewater treatment plant in South Korea: Guo Hong et al. proposed ANN and Support Vector Machine (SVM) models to predict the Total Nitrogen (T-N) concentration in the Yong-Yeon (YY) WWTP in Ulsan, South Korea [3]. For evaluation of the model, Coefficient of Determination (R^2), Nash-Sutcliff efficiency and relative efficiency criteria were used. A sensitivity analysis was done using a pattern search algorithm and Latin Hypercube One Factor at a Time (LH-OAT) [4] which showed that the ANN model gave superior results as compared to the SVM model. Resources used in the research are costly and it might not be possible to propose this model for the prediction of other effluent particles like TSS as the quality parameters are dependent on the site. In WWTPs at Korea, the anaerobic digestion process of sewage sludge with Food Waste has been increased. The increase of the Food Waste adversely affects digestion process and results in getting poor quality of effluent water from WWTPs.

3) Wastewater treatment plant in Italy: Granata Francesco et al. [5] conducted a study on stormwater discharge and proposed a model for the estimation of COD, BOD, TSS, and Total Dissolved Solids (TDS) in the wastewater. Support Vector Regression (SVR) and Regression Tree algorithms were used for modeling, and Coefficient of determination (R^2) and Root Mean Squared Error (RMSE) were the performance evaluators. For COD, TSS and TDS, the SVR model performed better than the Regression tree while for BOD, Regression Trees gave better results than SVR. Extending the proposed model to another treatment plant might not be a good choice as the conditions satisfying the development are heavily dependent on rainwater and only in a specific climate and rainy weather.

4) Wastewater treatment plant in Hong-Kong: Qin Xusong et al. [6], showed that wastewater quality can be monitored online. In this paper, UV/ VIS spectrometry and a turbid-meter were used to monitor COD, TSS, and Oil & Grease concentrations. Signals from the two sensors were fused using Sensor fusion technique. Boosting-Partial Least Squares (Boosting-PLS) [6] method was used to make the model and predict the wastewater quality based on the fused information.

5) Gongxian wastewater treatment plant in yibin, china: Wang Rui et al. [7] used four machine learning methods (Linear Regression, Ridge, Lasso and ElasticNet) for predicting the influent parameters. For influent parameter predictions, these methods showed high accuracy. The model proposed is used as warning module for assisting in daily operations of WWTP.

B. National Status

1) Wastewater treatment plant in Mangalore: D S Manu et al. used an artificial intelligence-based model [8] to predict the performance of the treatment plant for the removal of effluent nitrogen particles. Three different techniques/ models, SVM, ANFIS trapezoidal MF model and ANFIS Gbell MF model [8] were used and were implemented in MATLAB. Influent parameters taken were pH, ammonia nitrogen, free ammonia, and Kjeldahl nitrogen. Performance evaluation was done by RMSE, MSE, and Correlation Coefficient (R). The SVM model gave satisfactory results. This research is limited to a biological wastewater treatment plant and the study is conducted considering the biological waste only, thereby extending the proposed model to another site with a different type of wastes might not be possible.

2) Sewage treatment plant (STP) in Delhi: Gautam et al. [9] focused on the monitoring of inlet and outlet parameters and measuring the effectiveness of STPs in Delhi, India. The cluster analysis approach was performed to find any relation between the current site and other sites, aiming to find similar sites. Sulfate, Nitrates, Chloride and Phosphate, and Bi-carbonates concentrations were measured and the results showed that STP efficiency was not up to the mark. Samples were collected manually and the scope of automating the analyzing process is very limited. As the study is a decade old, the method used might be good for the estimation but the time consumed in the process can be reduced, if it were to be conducted and monitored online.

3) 345 MLD UASB-based Bharwara STP/ WWTP: Banerjee et al. [10] focused on the working performance of STP and upgrading Up-flow Anaerobic Sludge Blanket (UASB) reactor technology. The removal efficiency of COD, BOD, and TSS was measured and the relation between pH and influent parameters was determined. Measurement of the parameter was done two times in a month over a period of four months. Hence, this model is susceptible to climate changes, and a small amount of training data might not be a good choice for predicting the quality parameters.

III. Current Technological Status

As of now many researchers have contributed in proposing machine learning models [2, 3, 5, 6, 7, 8, 9, 10] for wastewater treatment plants. However, the available monitoring technologies used for analyzing and predicting parameters of wastewater quality have a number of limitations or drawbacks e.g., models are suited for plants of outside India [2, 3, 5, 6, 7]. In India, such smart systems for monitoring, analyzing, predicting quantity and quality parameters of wastewater are at a very preliminary stage. Smart, advanced and/ or automated systems are required not only for efficient utilization of resources but for smooth functioning of the wastewater treatment plants also, which can directly affect the health of humans/ living beings dependent on them. In India, no existing standard smart monitoring tool is currently available/ used [8, 9, 10] that predicts flow of influent and parameters of influent and effluent in advance for effective resource utilization at plant. The automation of system at WWTPs is still at an immature stage and not as developed as other process industries. The authors are researching in the line of implementing one such model/ infrastructure/ framework for wastewater treatment plant where data analysis shall be performed on real time data collected from Bharwara wastewater treatment plant, Lucknow (Fig. 1). In this paper, authors have presented the preliminary findings related to influent and effluent parameters of Bharwara wastewater treatment plant.

A. Pre-Treatment

The raw water is brought to the inlet chamber of Bharwara wastewater treatment plant (345 MLD UASB-based Bharwara STP) from the existing rising main of Gwari pumping station. The function of inlet chamber is to break the pressure flow and allow the sewage to flow by gravity to treatment unit. This chamber shall also function as flow distribution chamber to the screen channel.

Bharwara wastewater treatment plant is Asia's Largest UASB-based Wastewater Treatment Plant. The process flow of Bharwara wastewater treatment plant is shown in Fig. 2 and the brief description to its treatment scheme is given above.

Fig. 1. Aerial Photograph of Bharwara WWTP.

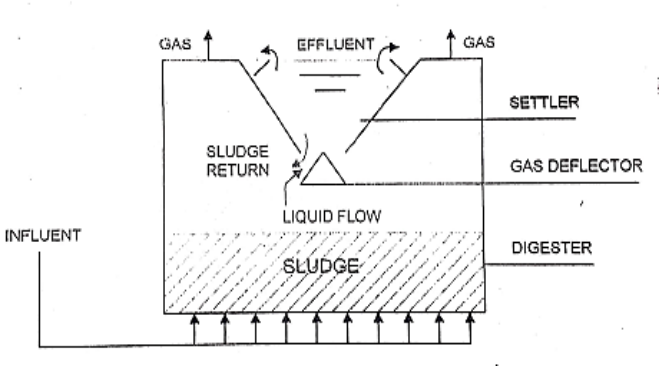


Fig. 2. Schematic of a UASB Reactor.

B. Primary Treatment

Primary treatment consists of Fine Screening and De-Gritting. Objective of Screening is to remove floating matters and other large size objects from the sewage stream. Objective of De-gritting is to remove grit particles from the sewage stream by gravity separation process. Screening and De-gritting are physical processes and they are accomplished in the units provided for the same with assistance of equipment provided. In the plant, nine automatic fine screening with 6 mm bar spacing are used. The cleaning of the screen is automated by mechanical means. Mechanical Fine Screen Channel always remains connected to the system, except during maintenance period. In the plant, three manual screening with a bar spacing of 12 mm (Standby for the automatic) are used. Six De-gritting Units with Grit Removal Mechanism and grit washing system are used for separation of gritty matter from sewage stream. Then, it is washed to make it free from organic matter and to transfer organic matter back to sewage stream. All the Grit Chambers are provided with Grit Collector, reciprocating type grit washing mechanism and Organic Return Pumps. These Grit Chambers remains in operation all the time. Three of them are manual De-gritting units for handling 50% of raw sewage flow and three of them are Parshall Flume with flow measurement followed by thirty Up-flow Anaerobic Sludge Blanket (UASB) Reactors as shown in Fig. 3. Inside the Reactors four reactions (Hydrolysis, Acidogenesis, Acetogenesis and Methanogenesis) account for the whole process. The system achieves a removal efficiency of 70%-80%, even when receiving organic loads greater than 15kg COD/m^3 of reactor per day at 8 hours HRT. The biogas is made up of 75%-85% methane. The sludge at the bottom has a concentration of about 40-70g Volatile Suspended Solids (VSS)/ l. In the plant, three Primary Sludge sump and pump house ultrasonic flow meters are provided in the Parshall flume for flow measurement with flow indicator, totalizer and recorder.

C. Secondary Treatment

In Secondary treatment, the refined water coming from primary treatment passes through three Pre-Aeration Tanks with six surface Aerators, further it has three Polishing ponds with two compartments having 18 Floating Aerators for polishing pond compartment 1 and eleven fountain pumps for polishing pond compartment 2.

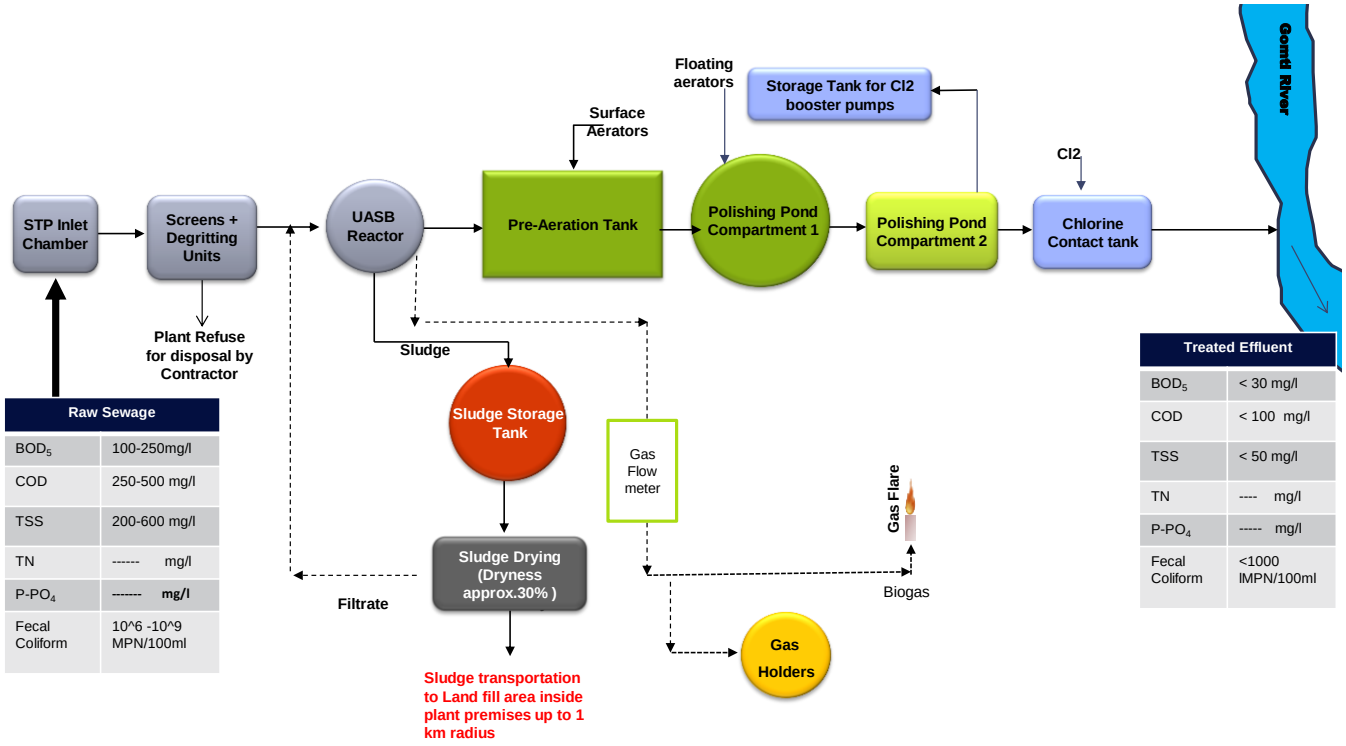


Fig. 3. Process Flow for 345 MLD UASB-based Bharwara STP.

D. Final Disinfection

In this process, three Chlorinator / one Chlorinator House is used. The chlorinator flow is manually adjusted. All the chlorinators are of the vacuum type. Hence, since as the chlorine gas is fed to the injector (located close to the chlorinators) at a pressure lower than atmospheric, no leak will occur. It has one water reservoir for chlorinating system and three Chlorine Contact Tank, One Final effluent chamber and One Final Effluent pipe.

The dewatering is carried out by means of Sludge drying beds. The 106 numbers of drying beds each of size 729 m² are feed by pumping from the SDB (Sludge Drying Bed) Feed sludge pumps. The consistency of dewatered sludge is around 30 – 35%.

The 345 MLD UASB-based Bharwara STP has following process objectives: reusable treated effluent, generating biogas according to raw sewage effluents and constantly delivering required quality of treated effluent.

In this section, we have briefly discussed major components of process flow for the plant.

In the next section, we present the methodology used to analyze and design a model to predict the parameters of influent/ effluent at wastewater treatment plant. This study shall provide foundation to improve reactor work performance in wastewater treatment plant.

IV. Methodology

We propose a machine learning based model to predict parameters of influent and effluent which shall provide efficient utilization of chemical resources during treatment process ensuring the desired level of quality indicators in effluent. We collected real time dataset of 345 MLD UASB-based Bharwara STP using manual process for data analysis. The methodology for the proposed model is briefed using the following four steps:

A. Identification of Locations and Water Parameters to be captured at Plant

We along with supporting staff at 345 MLD UASB-based Bharwara STP identified five locations where the water parameters are to be captured. The placing of various locations in the plant are shown in Fig. 2. At each location, we identified and listed the water parameters like BOD, COD, DO, SS, temperature, pH, Residual Chlorine etc. be measured. The basis of identifying water parameters at a particular location in the plant is the process/ treatment/ chemical reactions taking place at these locations. These identified locations and respective parameters to be measured at these locations are listed in Table I.

TABLE I. Locations and Measuring Parameters

Location	Parameters
Inlet Chamber	BOD, pH, Suspended Solids, Temperature, COD, oil, flow, Phosphorous, DO
Outlet of UASB Reactor	BOD, Suspended Solids, pH, COD
Polishing pond	Dissolved Oxygen, pH
Outlet of Chlorine contact Tank	BOD, Suspended solids, pH, COD, Fecal Coliform, Residual Chlorine, Dissolved Oxygen.
Primary sludge	pH, Total Solids, Volatile solids.

B. Data Collection

We collected a real-time data set of the 26 months (April 2019 to May 2021) from the plant. In the data set, selected parameters of influent and effluent are collected/ captured and recorded using manual process adopted at the plant.

C. Data Preprocessing

We do pre-processing on the recorded data set. For preprocessing, we treat missing values and outliers using standard procedures and kNN, and further normalized the data set. The outlier treatment is performed using statistical techniques i.e., calculating interquartile range and neglecting the values above lower limit and upper limit [11]. The normalization of the data set is performed using the following formula (1):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where x' is the normalized value, x is the original value, and min(x) and max(x) respectively are the minimum and maximum values. The data is normalized in the range between 0 and 1.

D. Discovering Unknown Patterns

We discover various patterns or relations within the collected data sets. We visualize the patterns in the data set. We design and implement a machine learning-based model to analyze and predict the parameters of influent/ effluent in the wastewater treatment Plant. We used Linear Regression to design the preliminary prediction model. Linear regression [12] is a statistical tool for the prediction of a dependent variable from an independent variable. It establishes a linear relationship between the independent (input) and dependent (output) variables. Linear Regression is a modeling technique where a dependent variable is predicted based on the independent variables. Linear Regression is the most widely used technique among all statistical techniques. The linear regression model is designed on Google Colab using python 3.7.12 for performing analysis.

Let us discuss the dependent variable, independent variable, line of regression, data preprocessing, model properties for the linear regression model.

Dependent variable: It is a variable that depends on other factors (independent variables) that are measured.

Independent variable: It is the variable [13] that is stable and unaffected by another variable which we are trying to measure independent variables (predictors) are used to predict the value of the dependent variable (target variable).

Line of regression model: It is the relationship between independent and dependent variables.

Model Properties: We implemented the initial model using Linear Regression in Python Implementation environment for the model is given in Table II.

TABLE II. IMPLEMENTATION ENVIRONMENT

Language	Python (version 3.7.12)
Tool	Google Colaboratory
Libraries	NumPy, Pandas, Matplotlib, Scikit Learn, SciPy and Seaborn

The model Properties are as follows:

- Model inputs: Inlet COD, BOD, PH, TSS, MLD and MPN
- Model outputs: Outlet COD, BOD, PH, TSS, DO, MPN
- Training - Test split: 80/20
- Estimator Function: Mean Square Error

In order to measure the performance of the model, Mean Square Error (MSE) is used. Formula for MSE [14] is given as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

We improved and tested the performance of the model by minimizing MSE and maximizing the correlation coefficient (R).

V. RESULTS AND ANALYSIS

We collected the raw data from 345 MLD UASB-based Bharwara STP during April 2019 to May 2021. The summary of the collected data, analysed using Python is shown in Table III. Based upon the initial analysis, it is found that the data has some missing facts/ details/ values and the outliers under few variables. Therefore, we applied Mean method and KNN to treat missing values in the given dataset. Fig. 4 shows the results after treating missing values on OUT_MPN. However, the similar results are obtained for the other variables (columns) with missing values in the dataset.

The raw data contained the outliers which were impacting/ decreasing the efficiency of the model(s) created. So, the treatment for removing the outliers from the data is done during pre-processing. The Fig. 5 shows the graph for inlet BOD before the outlier treatment and Fig. 6 shows the graph for inlet BOD after outliers' treatment. Similarly, we did the outlier treatment for other variables (columns) in the data set.

After doing the missing values treatment and outliers' treatment, the data set is used for further data visualization. We obtained the summary of the pre-processed dataset as shown in Table IV.

Further, we did data visualization and selected the influent parameters and obtained the results. The obtained results are elaborated in this section.

TABLE III. SUMMARY OF RAW DATA BEFORE PRE-PROCESSING

		DATA QUALITY PARAMETERS										
	Month Count	*MLD*	*Influent PH*	*Effluent PH*	*Influent TSS*	*Effluent TSS*	*Influent COD*	*Effluent COD*	*Influent BOD*	*Effluent BOD*	*Influent MPN*	*Effluent DO*
COUNT	781	780	780	780	780	780	780	780	774	774	572	780
MEAN	13.31	324.8	7.342	7.58	233.38	41.97	279.18	74.05	146.12	26.13	533836.14	4.58
STD	7.40	40.84	0.13	0.11	37.81	3.96	52.03	8.66	21.22	1.73	776153.06	1.41
MIN	1	157.11	6.93	7.23	145	4	125	28	75	19	1.4	4
25%	7	304.14	7.25	7.52	212	39	240	68	135	25	11	4.3
50%	13	344	7.34	7.6	230	42	264	72	148	26	780000	4.5
75%	20	349.79	7.43	7.66	254	45	324	80	160	28	920000	4.8
MAX	26	437.93	7.75	7.87	556	72	528	96	220	29	14000000	43

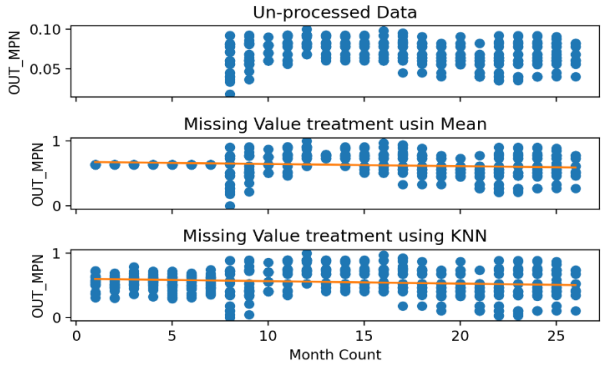


Fig. 4. Preprocessing of OUT_MPN.

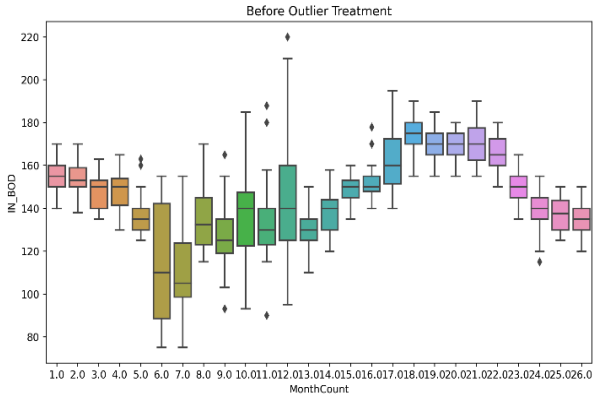


Fig. 5. Before Outlier Treatment.

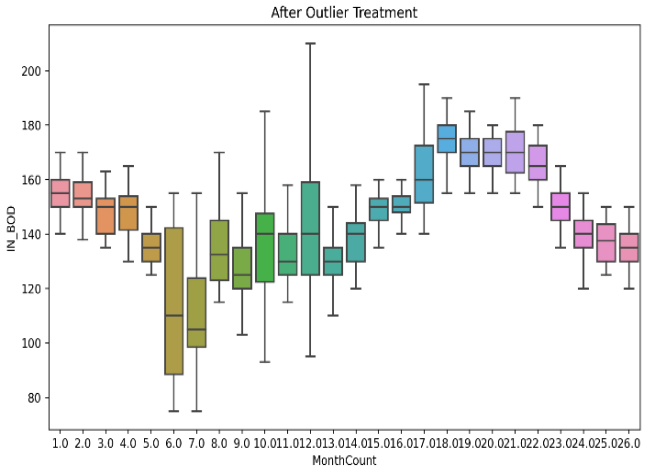


Fig. 6. After Outlier Treatment.

Fig. 7 shows the flow rate (in MLD) of influent with respect to month. In Fig. 8, we can clearly observe the inconsistency in the flow. Fig. 9 shows day wise flow of influent in the month of January of 2020 and 2021. For the month of February 2020, the flow is around the 7000 million litres but it rises too nearly 12000 million litres in the month of July. Also, the same months of different years have shown the major differences in the data which can be clearly observed in Fig. 9. We have further created the linear regression model for prediction of flow in 345 MLD UASB-based Bharwara STP/ WWTP.

TABLE IV. SUMMARY OF DATASET AFTER PRE-PROCESSING

		DATA QUALITY PARAMETERS											
	Month Count	*MLD*	*Influent PH*	*Effluent PH*	*Influent TSS*	*Effluent TSS*	*Influent COD*	*Effluent COD*	*Influent BOD*	*Effluent BOD*	*Influent MPN*	*Effluent MPN*	*Effluent DO*
COUNT	773	773	773	773	773	773	773	773	773	773	773	773	773
MEAN	13.23	326.2	7.34	7.58	231.51	42.01	276.81	74.01	145.77	26.14	10.31	0.06	4.53
STD	7.34	39.66	0.13	0.11	34.21	3.57	50.79	8.24	20.72	1.72	1.72	0.01	0.31
MIN	1	157.1	6.98	7.23	145	32	125	56	75	19	5.61	0.033	4
25%	7	306.8	7.25	7.52	212	39	240	68	135	25	9.2	0.061	4.3
50%	13	345.2	7.34	7.61	227	42	264	72	145	26	10	0.06	4.5
75%	20	350	7.43	7.66	251	45	320	80	160	28	11.21	0.08	4.8
MAX	26	403	7.75	7.87	443	49	464	96	210	29	15.51	0.1	5.3

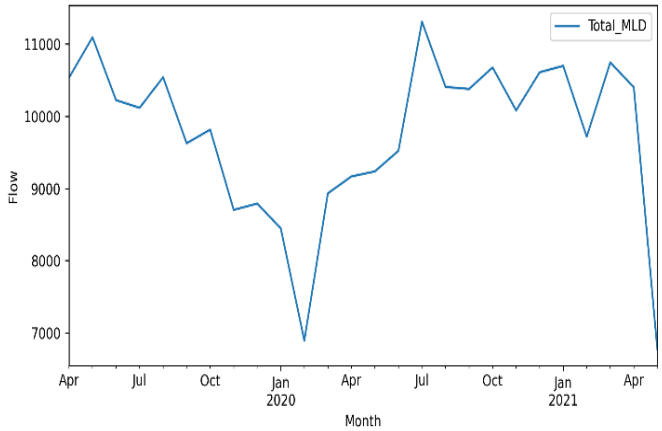


Fig. 7. Flow Rate of Influent by Month.

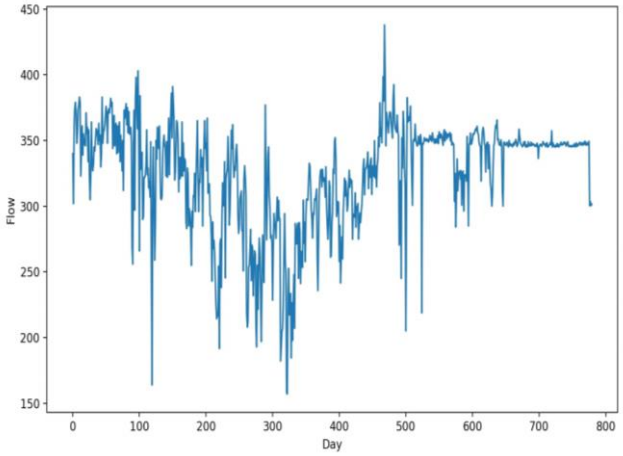


Fig. 8. Flow Rate of Influent by Day.

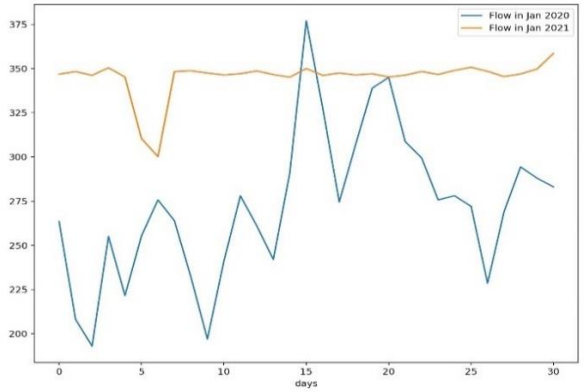


Fig. 9. Comparison of Flow Rate between Jan 2020 and Jan 2021 by Day.

Fig. 10 signifies the effectiveness of WWTP by showing the relationship between influent (untreated water) and Effluent (Treated and processed) water. The parameters include Total Suspended Solids (TSS), Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), Myeloproliferative Neoplasms (MPN), Dissolved Oxygen (DO) and pH.

These graphs in Fig. 10, show that there is a great fluctuation/ variation in the influent parameters which are the main factors affecting the efficiency of the plant. Therefore, a prediction by a Machine Learning model shall greatly help in managing and enhancing the quality and effectiveness of waste water treatment processes used in the plant.

The relationship between parameters can be analysed by the correlation coefficient. It can be used to obtain the effectiveness of the relationship among the parameters and can be used for further analysis and modelling. The positive correlation signifies that if one value increases another also increases, higher value shows the stronger correlation.

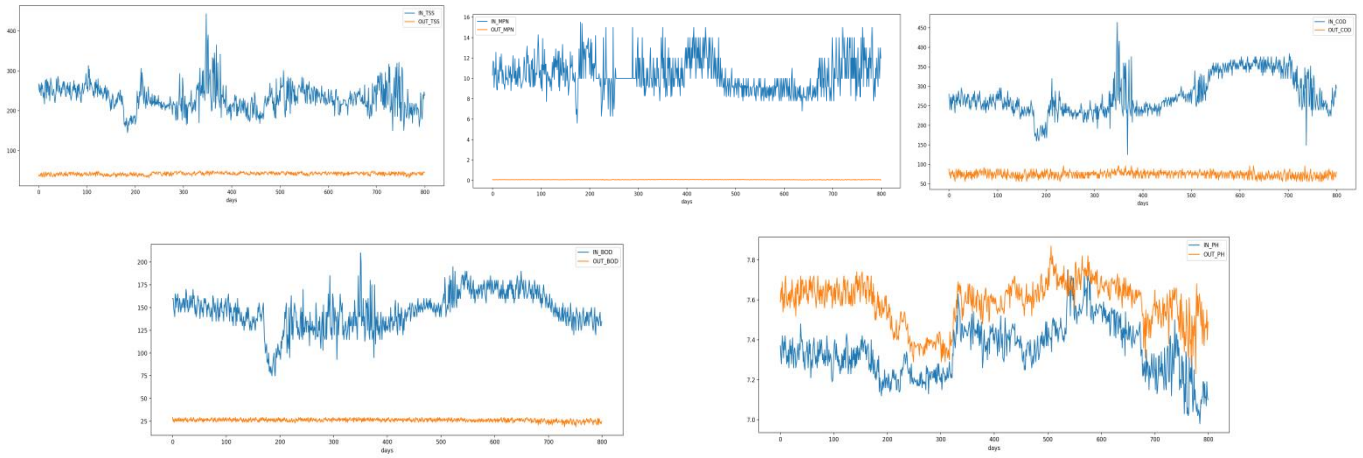


Fig. 10. Relationship between Influent and Effluent Parameters.

Fig. 11 shows the result of the linear regression model designed and implemented for prediction of effluent parameters (BOD, TSS, MPN, DO, and COD), here red dots show actual results and blue cross show predicted values of the linear regression model. The Model is trained on considering each influent parameter (BOD, TSS, MPN, DO, and COD) as input variables and the selected effluent parameter as output variable. The initial model is showing adequate results.

In Fig. 12, heatmap shows this relation with the intensity of the colour used; darker colour shows the stronger relationship. The colour turning to blue shows the negative relationship means the increase in one value will lead to the decrease in another.

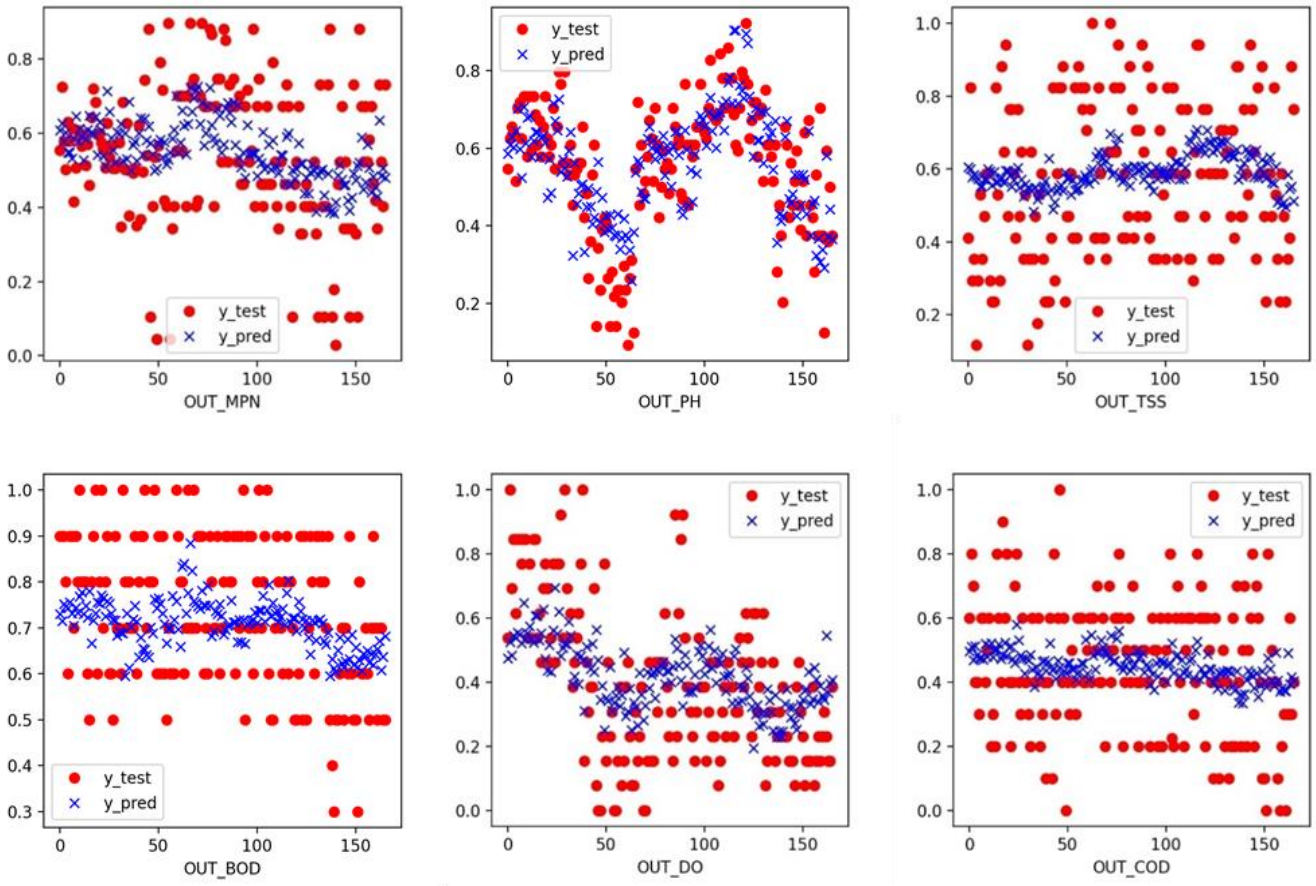


Fig. 11. Linear Regression Output.

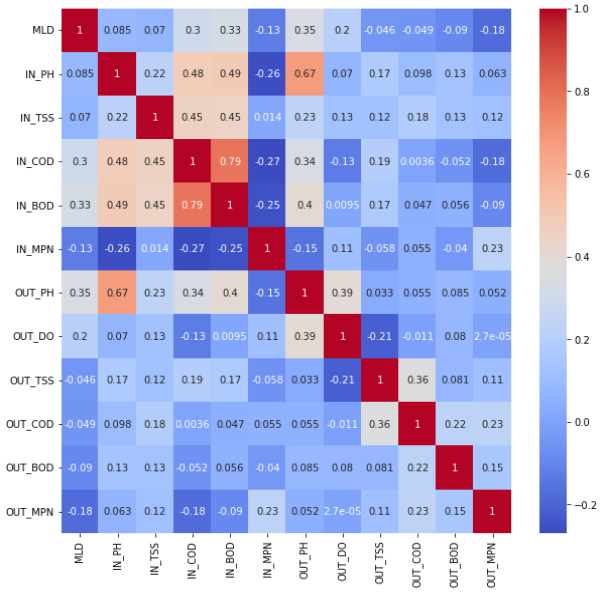


Fig. 12. Heatmap for Correlation.

Results of linear regression model, obtained in the form of MSE and Correlation Matrix (R), are shown in Table V. We are able to improve the efficiency of the initial model up to some extent using the linear regression model.

TABLE V. MSE TABLE AND CORRELATION

METHOD	DATA QUALITY PARAMETERS					
	EFFLUENT PH	EFFLUENT DO	EFFLUENT TSS	EFFLUENT COD	EFFLUENT BOD	EFFLUENT MPN
MSE	0.015	0.048	0.046	0.038	0.028	0.033
R	0.736	0.413	0.211	0.249	0.239	0.436

VI. CONCLUSION

Authors have analyzed the flow and quality parameters like COD, BOD, TSS, DO, pH, Temperature, Ammonia, Phosphorous and oil content, etc. in influent, and also parameters like COD, BOD, DO, pH, etc. of effluent in the WWTP. The proposed model provides support to centrally monitor processes and operations of a wastewater treatment plant. This paper depicts and visualizes the fluctuating and varying nature of influent parameters in 345 MLD UASB-based Bharwara STP.

The analysis presented in the paper, provides basis for improving operational efficiency and provides a cost-effective utilization of various resources at wastewater treatment plants by knowing about the pattern of the influent and effluent parameters may be in advance also. We have also designed and implemented an initial model using the Linear Regression algorithm to analyze as well as predict the parameters of influent and effluent of 345 MLD UASB-based Bharwara STP. However, the implemented model shall be applicable for any UASB based wastewater treatment plant or any wastewater treatment plant after a specific training part.

In future work, the authors shall incorporate SVM (Support Vector Machine) and ANN (Artificial Neural Network) techniques in the model to predict the influent and effluent parameters in WWTPs. It is expected that the model incorporating SVM and ANN shall show more robust relationship among the parameters and give a better estimate than the current model, in future.

REFERENCES

[1] Gautam Rajneesh, Verma Saumya, More Nandkishor. “Sewage Generation and Treatment Status for the Capital City of Uttar Pradesh, India”, Avicenna Journal of Environmental Health Engineering. vol. 5, no. 1, pp. 8-14, 2018.

[2] Tümer Abdullah, Edebali Serpil. (2015). An Artificial Neural Network Model for the Wastewater Treatment Plant of Konya. International Journal of Intelligent Systems and Applications in Engineering. vol. 3, no. 4, pp. 131-135, 2015.

[3] Guo Hong, Jeong Kwanho, Lim Jiyeon, Jo Jeongwon, Kim Young, Park Jong-pyo, Kim Joon Ha, Cho Kyung, “Prediction of effluent concentration in a wastewater treatment plant using machine learning models”, Journal of Environmental Sciences, vol. 32, pp. 90-101, 2015.

[4] Matala Anna., "Sample size requirement for Monte Carlo simulations using Latin hypercube sampling.", Helsinki University of Technology, Department of Engineering Physics and Mathematics, Systems Analysis Laboratory, 2008.

[5] Granata Francesco, Papirio Stefano, Esposito Giovanni, Gargano Rudy, DE MARINIS, Giovanni., “Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators”, Water, vol. 9, no. 2, pp. 105, 2017.

[6] Qin Xusong, Gao Furong, Chen Guohua, “Wastewater quality monitoring system using sensor fusion and machine learning techniques”, Water research, vol. 46, no. 4, pp. 1133-1144.

[7] Wang Rui, Pan Zhicheng, Chen Yangwu, Tan Zhouling, Zhang J. “Influent Quality and Quantity Prediction in Wastewater Treatment Plant: Model Construction and Evaluation”, Polish Journal of Environmental Studies, vol. 30, no. 5, 20.

[8] D S, Manu, Thalla Arun, “Artificial Intelligence Models for Predicting the Performance of Biological Wastewater Treatment Plant in the removal of Kjeldahl Nitrogen from Wastewater”, Applied Water Science, vol. 7, no. 7, pp. 3783-3791, 2017.

[9] Gautam, Sandeep Kumar, Divya Sharma, Jayant Kumar Tripathi, Saroj Ahirwar, and Sudhir Kumar Singh. "A study of the effectiveness of sewage treatment plants in Delhi region." Applied Water Science, vol. 3, no. 1, pp. 57-65, 2013.

[10] Banerjee, Arif Siddiquie1 Rajiv. "Performance Evaluation & Upgradation of UASB Technology used for the Treatment of Sewage Generated from Lucknow City.", 2016.

[11] Alnaa, Samuel, Ahiakpor, Ferdinand, “ARIMA (autoregressive integrated moving average) approach to predicting inflation in Ghana”, Journal of Economics and International Finance, vol.3, pp. 328-336, 2011.

[12] Kumari, Khushbu, and Suniti Yadav. "Linear regression analysis study." Journal of the practice of Cardiovascular Sciences, vol. 4, no. 1, pp. 33, 2018.

[13] Cramer, Duncan, and Dennis Laurence Howitt. The Sage dictionary of statistics: a practical resource for students in the social sciences, Sage, 2004.

[14] Available at: https://www.datavedas.com/model-evaluation-regression-models/.

[15] Sin, Gürkan, Krist V. Gernaey, Marc B. Neumann, Mark CM van Loosdrecht, and Willi Gujer. "Uncertainty analysis in WWTP model applications: a critical discussion using an example from design.", Water Research, vol.43, no. 11, pp. 2894-2906, 2009.

Forecasting Foreign Currency Exchange Rate using Convolutional Neural Network

Manaswinee Madhumita Panda[1]
Surya Narayan Panda[2]
Institute of Engineering and Technology
Chitkara University, Punjab, India

Prasant Kumar Pattnaik[3]
School of Computer Engineering
KIIT, Deemed to be University
Bhubaneswar, India

***Abstract*—Foreign exchange rate forecasting has always been in demand because it is critical for foreign traders to know how their money will perform against other currencies. Traders and investors are always looking for fresh ways to outperform the market and make more money. As a result, economists, researchers and investors have done a number of studies in order to forecast trends and facts that influence the rise or fall of the exchange rate (ER). In this paper, a new Convolutional Neural Network (CNN) model with a random forest regression layer is used for future closing price prediction. The intended model has been tested using three major currency pairs: Australian Dollar against the Japanese Yen (AUD/JPY), the New Zealand Dollar against the US Dollar (NZD/USD) and the British Pound Sterling against the Japanese Yen (GBP/JPY). As a proof-of-concept, the forecast is made for 1 month, 2 months, 3 months, 4 months, 5 months, 6 months and 7 months utilizing data from January 2, 2001 to May 31, 2020 for AUD/JPY and GBP/JPY and data from January 1, 2003 to May 31, 2020 for NZD/USD. Furthermore, when compared the performance of the suggested model with the Autoregressive Integrated Moving Average (ARIMA), Multi-Layer Perceptron (MLP) and Linear Regression (LR) models and found that the proposed CNN with Random Forest model surpasses all models. The suggested model's prediction performance is assessed using R^2, MAE, RMSE performance measures. The proposed model's average R^2 values for three currency pairs from one to seven months are 0.9616, 0.9640 and 0.9620, demonstrating that it is the best model among them. The study's findings have ramifications for both policymakers and investors in the foreign exchange market.**

***Keywords*—Convolutional neural network; exchange rate; R square; random forest regression method**

I. Introduction

In our lives, money is quite important. Currency markets have evolved into an important part of our lives in the growth of the financial system. The exchange rate (ER) and foreign exchange rate (Forex) are two well-known terminologies in the money market. ER is the worth of a country's currency that may be exchanged for another country's currency [1]. Forex is a global market where national currencies are exchanged for one another. Forex is the world's largest daily market for swapping one currency for another [2]. The ER is influenced by a variety of factors, including individual traders' and investors' economic, political and psychological circumstances [3,4]. The stunning presentation of the CNN on the detailed earth has gotten a lot of attention. Visual Speech recognition [5], Hand Gesture recognition [6], COVID-19 Detection [7], and time series data prediction [8] are all domains where CNN technology is used. CNN was once utilized to forecast stock market activity for the next day [9].

The primary goal of this study is to verify the computationally efficient Convolutional Neural Network model for forecasting foreign currency exchange rates. The suggested model (CNN-RF) was used to forecast the closing price of three key foreign currency pairs: the AUD/JPY, NZD/USD and GBP/JPY. The proposed model is used to forecast currency prices up to seven months ahead of time. The experimental findings are analyzed using three commonly used performance metrics: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAE) and the coefficient of determination (R^2). The suggested model is compared to ARIMA, MLP and LR approaches based on performance metrics. Although many researchers have been anticipating foreign exchange currency in the past, but currently researchers are striving to come up with new models to predict the character of this market. While there are numerous machine learning and deep learning approaches utilized in finance, traders are always looking for new ways to outperform the market. This model will assist traders in achieving their goals in a systematic manner.

The following are the study's primary research contributions: (1) To capture exchange rate uncertainty, a reliable, efficient and accurate forecasting model employing CNN with Random Forest regression layer is proposed and implemented. Every 1 minute opening price, closing price, high value, low value and volume of exchange rates are all taken into account. (2) Accuracy measuring methods such as MAE, RMSE, and R^2 are used to compare the results. The R^2 value is used to measure the system's performance. The proposed forecasting model is, to our knowledge, the first of its sort in the literature.

The organization of the paper is given as: The associated work for currency ER prediction is explained in Section II. We present the data used for doing the experiment & some preferred models including our suggested model used for prediction of ER in Section III. Performance evaluation criteria described in Section IV. Experimental outcomes discussed in Section V followed by conclusions in the next section.

II. Literature Survey

Countless studies have proposed and developed numerous ways to examine and predict ER activity in the last few years. The following is a concise discussion of the important investigations.

For time series forecasting a broad range of prediction methods have been measured. The ARIMA model was known as the Box-Jenkins model and affirmed that it is the most popular scheme used for time series forecasting [10]. The ARIMA-GARCH model is used to forecast Ghana's GDP. It demonstrates that the ARIMA-GARCH model may reduce error variance and improve forecasts [11]. In real-world monetary time series, the performance of the linear models is below expectation due to boundaries in linear models. Thus, in this document, we talk about various non-linear models like artificial neural networks. For non-linear prediction the non-linear model artificial neural network (ANN) is used [12].

Many academics have tried using the Long Short-Term Memory (LSTM) technique to anticipate currency exchange rates in recent years. LSTM networks operate well with time series data for classification, analysis and prediction. For short-term prediction, Elman type approaches such as ARIMA, LSTM and Recurrent Neural Network (RNN) are used. 5 days, 11 days, 22 days, 35 days, 44 days and 55 days windows were forecasted using the three approaches above. The 22-day window has an average accuracy of 71.76 percent. In the short term, the validation dataset was best approximated using the 22-day predicting window [13]. As forex market is very volatile & complex. Thus, all the investors find new methods with more accuracy. The suggested model mentioned in this paper gives 93% average accuracy for 1 month ahead prediction which is also considered as a short-term prediction. For projecting the price 10 and 30 minutes in advance, the Gated Recurrent Unit (GRU)-LSTM approach is used. The key currency pairs EUR/USD, GBP/USD, USD/CAD and USD/CHF were evaluated for this experiment. The performance of the GRU-LSTM model is compared to that of the GRU, LSTM and statistical models based on simple moving average (SMA). Based on MAE, MSE, and RMSE results the suggested model provides better results of GBP/USD and USD/CAD currency pair. The GRU-LSTM model outperforms other models in terms of R^2 [14]. Based on the evaluation criteria our proposed model CNN with Random Forest (CNN-RF) provides better results in all datasets. Three alternative models such as support vector regression, back propagation neural network and long short-term memory, were used using Google trends and macroeconomic data to predict the value of Ghanaian Cedis in USD, British Pounds and Euros for the next 30 days. The results reveal that, unlike the other two models, the LSTM can easily manage exchange rate data variance [15]. However, in any circumstance, it's possible that Google Trends won't be able to predict changes in the Ghanaian cedi's exchange rate versus all other currencies.

Another hybrid model, ANN-GJR (Glosten, Jagannathan and Runkle) was employed to forecast currency exchange rates using five currency pairs. When compared to the benchmark model, the ANN hybrid model performs better. 14 days, 21 days and 28 days are the predicting horizons. The hybrid model's forecasting precision is over 90% for a 21-day horizon. The hybrid model's prediction accuracy improves as the prediction horizon lengthens and the benchmark model performs better for shorter horizons [16]. Therefore, this model only applied for long term forecasting. The average accuracy of the proposed model (CNN-RF) is 93% for one month prediction and 95% accuracy for seven months ahead prediction. Thus, the proposed model performance is better in both a short and long time period. Alternative hybrid model ANN-GA (genetic algorithm) used for INR (Indian Rupees) Vs USD currency exchange rate prediction. But ANN has some limitations: requires a large diversity of training for operation and also have overfitting problem [17].

Support vector regression (SVR) method is used for forecasting short-term financial time series. The forecasts were produced one to four days in advance, with a focus on the short term. The proposed PCA-ICA-SVR model facilitates to forecast stock values with small amount error [18]. This method was also used by other researchers for currency exchange rate prediction [19].

SVR, NN, LSTM known as Support vector regression, neural network, long short-term memory with hidden layers used in deep learning models for multi-currency ER prediction. They forecast the ER between the top currencies of the world. The average precision of the forecasting model exceeds 99% [20]. Based on the outcome of the forecasting model the average accuracy is very good but it will predict data by day. There is no provision for long-term prediction. A new method deep belief network (DBN) used for forecasting the currency exchange rate data. For doing the experiment INR/USD and CNY/USD currency pair used. On the overall prediction performance, with the increase of the forecast period, the prediction accuracy declines. Thus, this model is not suitable for long term prediction [21].

III. Data and Models Apply for Forecasting Foreign Currency Exchange Rates

The data collection sources are mentioned in this section. Different algorithms for Forex are discussed in this section.

A. Data

The experiment was conducted using three major currency pairs: AUD/JPY, NZD/USD and GBP/JPY. January 2, 2001 to May 31, 2020 for AUD/JPY and GBP/JPY and data from January 1, 2003 to May 31, 2020 for NZD/USD data was collected for doing experiments. Everyday opening price, closing price, low price, high price and volume data were gathered. (http://www.forextester.com/data/datasources) site used for collecting past 1-minute trade rates information. To abolish unnecessary information, 1-minute data was resampled to 1hour, 1day data.

B. Auto Regressive Integrated Moving Average (ARIMA)

It is the most popular approach used for forecasting. The future value of a variable is a linear combination of past error and past value.

$$x_t = \alpha_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots\ldots\ldots + \beta_u y_{t-u} + \delta_t$$
$$-\alpha_1 \delta_{t-1} - \alpha_2 \delta_{t-1} - \ldots. - \alpha_v \delta_{t-v} \quad (1)$$

Where x_t represents actual value, δ_t represent random error at t time, α_i and β_j represents the coefficients, u and v integers frequently referred to as autoregressive and moving average polynomials [22]. ARIMA model consists of three phases: identification of model, estimation of parameters & diagnostic checking.

C. Multilayer Perceptron (MLP)

MLP is also used for ER prediction [23]. An MLP network has three types of layers: a hidden layer exists within the input and output layer. The hidden layer neuron adds the received input signal after multiplying all input signals to their associated weight values. Based on the received output of the neuron is estimated.

D. Random Forest (RF)

Random forest ensemble learning procedure applied for regression and classification problems. For foreign currency ERs prediction, we use the random forest method. It simply works on tabular data. The missing value will not be measured while we prepare the data in python and sklearn packages. The dataset, depth of tree and no of the tree are the inputs of the random forest (RF) method [24].

E. Linear Regression (LR)

Here for forecasting foreign currency ER we use the LR on technique. It is of two type's linear regression and multiple regression. The mathematical formula of linear regression is

$$w = by + c \tag{2}$$

W stands for dependent variable, y stands for the independent variable, b known as slope, c considers as interceptor.

$$w = by1 + by2 + by3 + c \tag{3}$$

w stands for the dependent variable. The independent variables are represented as y1, y2 and y3.b is the slope, c known as the interceptor [25].

F. CNN

In CNN several hidden layers exist within input and output layer. The hidden part of CNN is the combination of a convolution layer (CL), pooling layer and a fully connected (FC) classifier. The features are obtained from the previous layer processed by the fully connected classifier [26].

Fig. 1 shows the workflow diagram, which provides the framework of the proposed ER prediction model. The framework consists of some phases: (i) Data download and integration; (ii) Data pre-processing and partition; (iii) Convolution layer using leakyrelu activation function; (iv) Pooling layer; (v) Sigmoid mapping function used for selecting abstract feature; (v) Decision tree regression layer; (vi) Random forest regression layer; (vii) Analysis model performance; and (viii) Output.

For extracting local features from input a set of filters present in the CL. A very small part of the input is called the receptive field which represents feature map. The neuron present in the CL is very much associated with the tiny piece of the preceding layer. These neurons prepare the feature map and the similar weight pooled to every feature map. The pooling layer applied to shrink the extent of each feature map. Two types of pooling are used mean-pooling and max-pooling. Max-pooling is used for selecting the highest value and mean-pooling use for selecting the least amount of feature map.

The proposed model's architecture is depicted in Fig. 2. Our model has four Convolutional layers, three pooling layers and one fully connected layer & a readout layer. First, the inputs are pre-processed to a standard normalization (0,1). Then the feature and labels are selected. After that represent the feature matrix with label. PCA algorithm is used for data transformation or dimension reduction. Every minute of 24 hours data is taken as input. The max pooling layer is followed by the convolution layer. Finally, as a readout layer, a random forest regression layer is employed. To reduce the RMSE loss, we used Adam optimization to train the network. A batch size of 256 was chosen. The probability of dropping out rate is 20%. We trained for 10 epochs with a learning rate of 0.003 at first. We divide the complete data into two parts: 80 percent for training and 20 percent for testing.

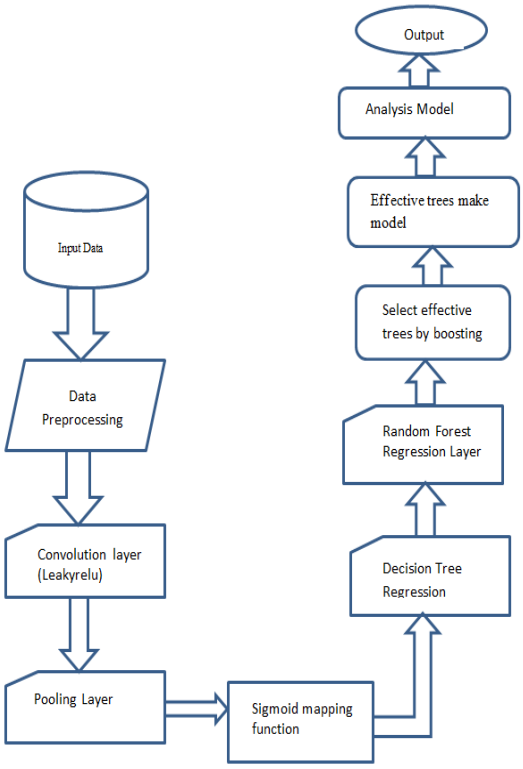


Fig. 1. Work Flow Diagram of Proposed Model.

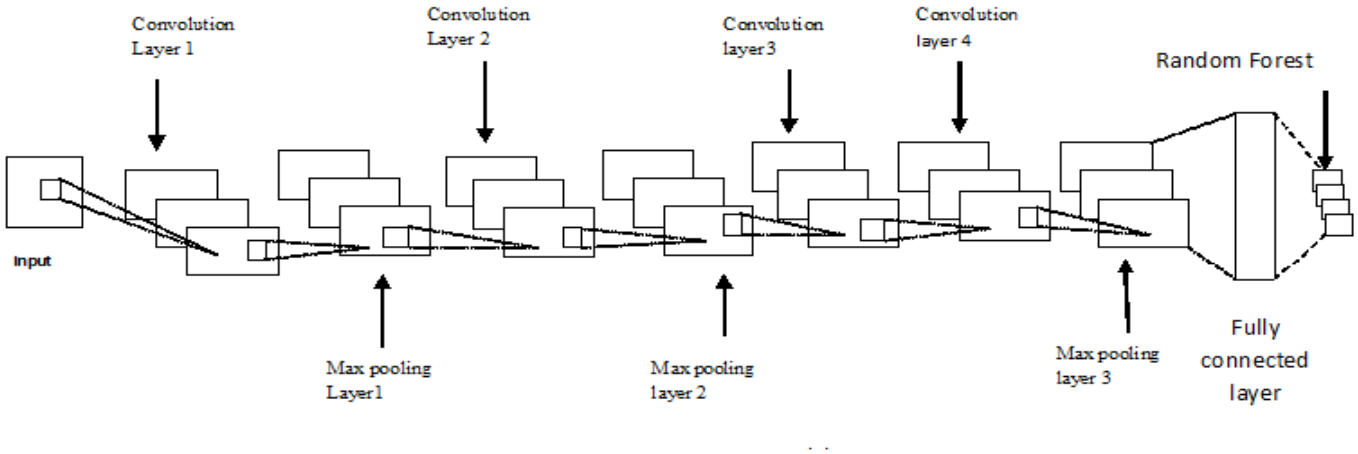


Fig. 2. Architecture of Proposed Method.

IV. Performance Evaluation Criteria

The model performance is measured by four indices R^2, RMSE, MAE. The MAE is computed as

$$MAE(u,v) = \frac{1}{n}\sum_{i=1}^{n}|u_i - v_i| \tag{4}$$

Where u_i is the real value of the i^{th} sample, v_i represents the expected value of i^{th} sample and n represents total number of samples. MAE calculates the distinction between real value and predicted value by average the absolute difference over the test sample

$$RMSE(u,\ v) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - v_i)} \tag{5}$$

RMSE is very much useful when the performance of the model is pretentious due to the presence of a large number of errors. The square root of the mean of the difference among the real and the expected values are considered.

$$R^2(u,v) = 1 - \frac{\sum_{i=1}^{n}(u_i - v_i)^2}{\sum_{i=1}^{n}(u_i - \overline{u_i})^2} \tag{6}$$

The actual value represented as $\overline{u_i}$ .To know the fitness of a forecasting model R^2 is used. The range of R^2 lies within 0 to 1 if the value is superior the model is better.

V. Results and Discussion

The intended model has been doing the experiment using three major currency pairs AUD/JPY, NZD/USD, GBP/JPY. Moreover, ARIMA, MLP, LR algorithms are considered for comparison with our proposed (CNN-RF) scheme. From the entire data, 80% and 20% of data is used for training and testing the suggested model. The proposed model is used to forecast up to 7 months in advance. The model's performance is calculated through R^2, MAE, and RMSE. The model has been designed & tested using PYTHON3.7 software.

A. AUD/JPY

From Table I, it can be observed that our suggested model is superior to other models. It is indicated in Table I, for the first month the R^2 (the greater value is superior) value is 46.92%, 70.39% and 17.37% higher than ARIMA, MLP and LR techniques. In addition to this, for a second month it also shows better R^2 performance compared to other existing models. Its performance for R^2, is higher than ARIMA, MLP and LR by 47.86%, 70.84% and 18.38% correspondingly. For the third month, the result of R^2 of the suggested model is 48.01%, 63.93%, 18.42% bigger than ARIMA, MLP and LR methods. For the fourth month, the result of R^2 of the suggested model is 49.53%, 60.76%, 20.26% bigger than ARIMA, MLP and LR techniques. For the fifth month, the result of R2 of the proposed model is 49.12%, 55.45%, 19.83% bigger than ARIMA, MLP and LR algorithms. For the sixth month the result of R^2 of the suggested model is 50.19%, 58.57%, 21.09% bigger than ARIMA, MLP and LR methods. For the seventh month, the result of R^2 of the suggested model is 50.34%, 57.53%, 20.45% bigger than ARIMA, MLP and LR methods.

The R^2 result of the AUD/JPY from 1 to 7 months forecast in advance is presented in Fig. 3 using several methodologies such as CNN-RF, ARIMA, MLP, and LR.

TABLE I. Comparison of R^2 for AUD to JPY Forex Data using Different Techniques

R^2(AUD/JPY)				
Months	CNN- RF	ARIMA	MLP	LR
1	0.9343	0.6359	0.5483	0.7960
2	0.9400	0.6357	0.5502	0.7940
3	0.9418	0.6363	0.5745	0.7953
4	0.9530	0.6373	0.5928	0.7924
5	0.9526	0.6388	0.6128	0.7949
6	0.9581	0.6379	0.6042	0.7912
7	0.9616	0.6396	0.6104	0.7983

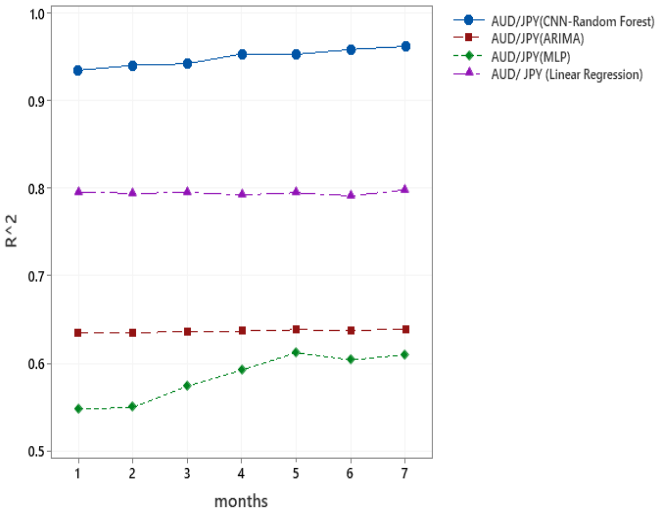


Fig. 3. ER of AUD / JPY Seven Months ahead Prediction Error (R^2).

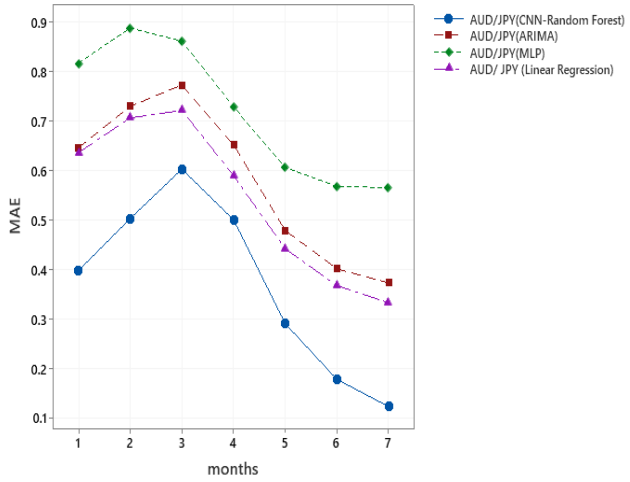


Fig. 4. ER of AUD / JPY Seven Months ahead Prediction Error (MAE).

It is shown in Table II that our suggested scheme performs superior to the state-of-the-art methods. For MAE (the least value is the better) performance of the proposed approach is better than all other models. When MAE performance measured in AUD/JPY exchange rate forecasting for the first month our suggested model result is 38.39%, 51.27%, 37.42% smaller than ARIMA, MLP and LR algorithms. For the second month, the result of MAE of the proposed model is 31.05%, 43.34%, 28.76% smaller than ARIMA, MLP and LR algorithms. For the third month, the result of MAE of the suggested model is 21.98%, 29.97%, 16.42% smaller than ARIMA, MLP and LR techniques. For the fourth month, the result of MAE of the proposed method is 23.24%, 31.29%, 15.07% smaller than ARIMA, MLP and LR algorithms. For the fifth month, the result of the MAE of the suggested model is 39.29%, 52.03%, 34.23% smaller than ARIMA, MLP and LR methods. For the sixth month, the result of MAE of the proposed model is 55.58%, 68.61%, 51.49% smaller than ARIMA, MLP and LR techniques. For seven month the result of MAE of the suggested model is 66.95 %, 78.19%, 62.98% smaller than ARIMA, MLP and LR methods.

Using multiple techniques such as CNN-RF, ARIMA, MLP, and LR, the MAE result of the AUD/JPY from 1 to 7 months forecast in advance is provided in Fig. 4.

TABLE II. COMPARISON OF MAE FOR AUD TO JPY FOREX DATA USING DIFFERENT TECHNIQUES

MAE(AUD/JPY)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.3984	0.6467	0.8177	0.6367
2	0.5034	0.7301	0.8886	0.7067
3	0.6034	0.7734	0.8617	0.7220
4	0.5009	0.6526	0.7291	0.5898
5	0.2909	0.4792	0.6065	0.4423
6	0.1784	0.4017	0.5684	0.3678
7	0.1234	0.3734	0.5659	0.3334

Table III represents the proposed model performs better than LR, ARIMA and MLP models for prediction of seven months. When RMSE performance measured in AUD/JPY ER forecasting for the first month our proposed model result is 26.25% smaller than ARIMA, 36.38% smaller than MLP, 21.75% smaller than LR. For the second month, the result of RMSE of the suggested model is 27.60% smaller than ARIMA, 38.01% smaller than MLP, 21.99% smaller than LR. For the third month, the result of RMSE of the proposed model is 34.90% smaller than ARIMA, 47.12% smaller than MLP, 29.27% smaller than LR. For the fourth month, the result of RMSE of the suggested model is 45.76% smaller than ARIMA, 59.21% smaller than MLP, 40.78% smaller than LR. For the fifth month, the result of RMSE of the proposed model is 55.61% smaller than ARIMA, 68.74% smaller than MLP, 51.25% smaller than LR. For the sixth month, the result of RMSE of the suggested model is 59.15% smaller than ARIMA, 71.87% smaller than MLP, 54.93% smaller than LR. For seven months the result of RMSE of the proposed model is 61.85% smaller than ARIMA, 74.15% smaller than MLP, 57.66% smaller than LR.

Using multiple techniques such as CNN-RF, ARIMA, MLP, and LR, the RMSE result of the AUD/JPY from one to seven months ahead forecast is shown in Fig. 5.

TABLE III. COMPARISON OF RMSE FOR AUD TO JPY FOREX DATA USING DIFFERENT TECHNIQUES

RMSE(AUD/JPY)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.5317	0.7210	0.8358	0.6795
2	0.4723	0.6524	0.7620	0.6055
3	0.3595	0.5523	0.6799	0.5083
4	0.2548	0.4698	0.6248	0.4303
5	0.1881	0.4238	0.6019	0.3859
6	0.1680	0.4113	0.5973	0.3728
7	0.1542	0.4042	0.5967	0.3642

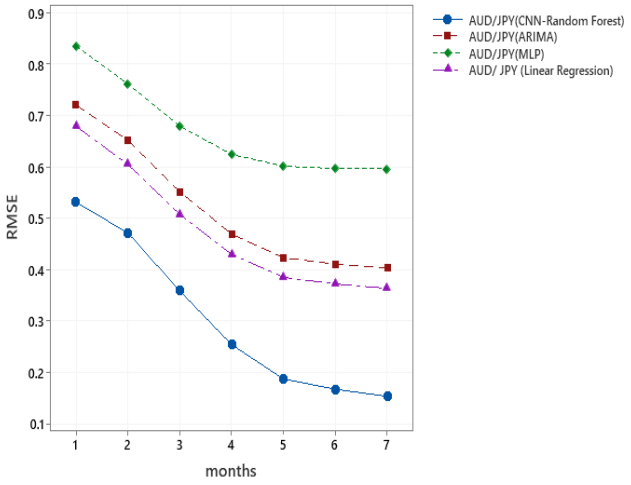


Fig. 5. ER of AUD / JPY Seven Months ahead Prediction Error (RMSE).

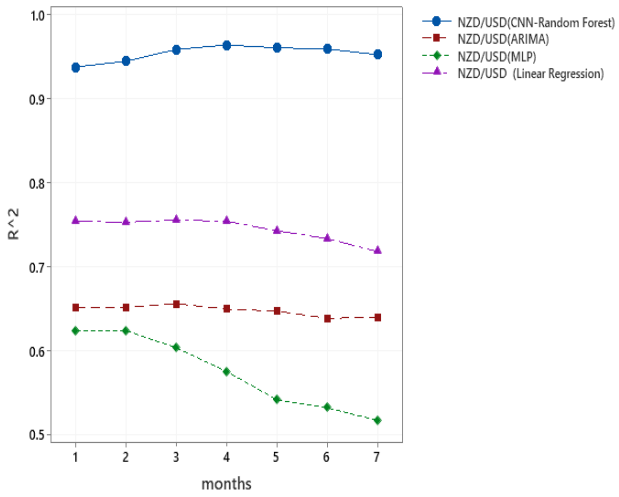


Fig. 6. ER of NZD / USD Seven Months ahead Prediction Error (R^2).

B. NZD / USD

In Table IV, the performance of the suggested model is evaluated. For the first month, the R^2 performance in NZD/USD ER, our proposed model result is 44.01% bigger than ARIMA, 50.40%bigger than MLP,24.36%bigger than LR. For the second month, the result of R^2 of the suggested model is 44.97% bigger than ARIMA,51.41% bigger than MLP,25.45% bigger than LR. For the third month, the result of R^2 of the suggested model is 46.13% bigger than ARIMA, 58.84% bigger than MLP, 26.89% bigger than LR. For the fourth month, the result of R^2 of the suggested model is 48.35% bigger than ARIMA, 67.56% bigger than MLP, 27.81% bigger than LR. For the fifth month, the result of R^2 of the suggested model is 48.43% bigger than ARIMA,77.31% bigger than MLP,29.33% bigger than LR. For the sixth month, the result of R^2 of the suggested model is 50.19% bigger than ARIMA, 80.25% bigger than MLP, 30.76% bigger than LR. For the seventh month, the result of R^2 of the suggested model is 48.85% bigger than ARIMA, 84.20% bigger than MLP, 32.66% bigger than LR.

The R^2 result of the NZD/USD from 1 to 7 months forecast in advance is shown in Fig. 6 using several methodologies such as CNN-RF, ARIMA, MLP, and LR.

TABLE IV. COMPARISON OF R^2 FOR NZD TO USD FOREX DATA USING DIFFERENT TECHNIQUES

R^2(NZD/USD)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.9378	0.6512	0.6235	0.7541
2	0.9448	0.6517	0.6240	0.7531
3	0.9588	0.6561	0.6036	0.7556
4	0.9640	0.6498	0.5753	0.7542
5	0.9607	0.6472	0.5418	0.7428
6	0.9593	0.6387	0.5322	0.7336
7	0.9525	0.6399	0.5171	0.7180

TABLE V. COMPARISON OF MAE FOR NZD TO USD FOREX DATA USING DIFFERENT TECHNIQUES

MAE(NZD/USD)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.3784	0.6217	0.8675	0.6465
2	0.4734	0.7134	0.9251	0.7064
3	0.5434	0.7501	0.8990	0.7051
4	0.4434	0.6067	0.7567	0.5744
5	0.2634	0.4567	0.6465	0.4496
6	0.1684	0.4017	0.5997	0.4106
7	0.1334	0.3834	0.5759	0.3959

Table V show that our suggested model works better than MLP, ARIMA and LR models. For MAE, the proposed technique shows remarkable performance than other existing methods. When MAE performance measured in NZD/USD ER forecasting for the first month, our proposed model result is 39.13% smaller than ARIMA, 56.38% smaller than MLP, 41.46% smaller than LR. For the second month, the result of MAE of the suggested model is 33.64% smaller than ARIMA, 48.82% smaller than MLP, 32.98% smaller than LR. For the third month, the result of MAE of the proposed model is 27.55% smaller than ARIMA, 39.55% smaller than MLP, 22.80% smaller than LR. For the fourth month, the result of MAE of the suggested model is 26.91% smaller than ARIMA, 41.40% smaller than MLP, 22.80% smaller than LR. For the fifth month, the result of MAE of the proposed model is 42.32% smaller than ARIMA, 59.25% smaller than MLP, 41.41% smaller than LR. For the sixth month, the result of MAE of the suggested model is 58.07% smaller than ARIMA, 71.91% smaller than MLP, 58.98% smaller than LR. For seven months, the result of MAE of the proposed model is 65.20% smaller than ARIMA, 76.83% smaller than MLP, 66.30% smaller than LR.

Using multiple techniques such as CNN-RF, ARIMA, MLP, and LR, the MAE result of the NZD/USD from 1 to 7 months ahead is displayed in Fig. 7.

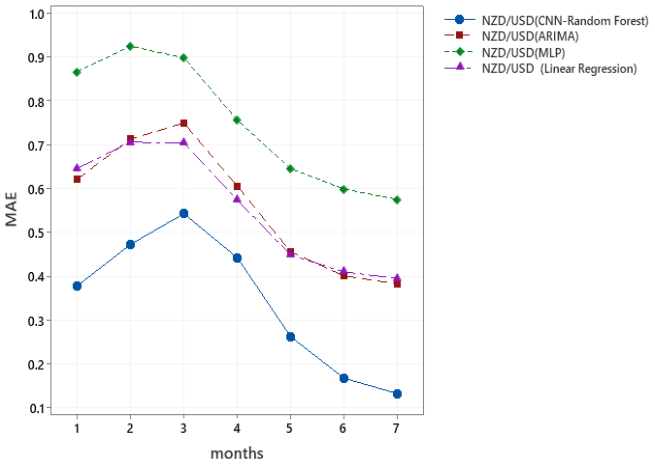


Fig. 7. ER of NZD / USD Seven Months ahead Prediction Error (MAE).

TABLE VI. COMPARISON OF RMSE FOR NZD TO USD FOREX DATA USING DIFFERENT TECHNIQUES

RMSE(NZD/USD)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.4870	0.6940	0.8724	0.6722
2	0.4292	0.6251	0.7959	0.6030
3	0.3298	0.5331	0.7116	0.5219
4	0.2411	0.4624	0.6517	0.4642
5	0.1884	0.4274	0.6212	0.4367
6	0.1730	0.4188	0.6127	0.4304
7	0.1642	0.4142	0.6067	0.4267

In Table VI, it is observed that the suggested model shows better than LR, ARIMA and MLP models. For the RMSE parameter (the least value is the better), the performance of the suggested scheme is better than other models . In first month, the RMSE parameters in NZD/USD ER by our proposed model result is 29.82% smaller than ARIMA,44.17% smaller than MLP,27.55% smaller than LR. For the second month the result of RMSE of the suggested model is 31.33% smaller than ARIMA, 46.07% smaller than MLP, 28.82% smaller than LR. For the third month the result of RMSE of the proposed model is 38.13% smaller than ARIMA, 53.65% smaller than MLP, 36.80% smaller than LR. For the fourth month the result of RMSE of the suggested model is 47.85% smaller than ARIMA, 63.00% smaller than MLP, 48.06% smaller than LR. For the fifth month the result of RMSE of the proposed model is 55.91% smaller than ARIMA, 69.67% smaller than MLP, 56.85% smaller than LR. For the sixth month the result of RMSE of the suggested model is 58.69% smaller than ARIMA, 71.76% smaller than MLP, 59.80% smaller than LR. For seven month the result of RMSE of the proposed model is 60.35% smaller than ARIMA, 72.93% smaller than MLP, 61.51% smaller than LR.

The RMSE outcome of the NZD/USD from 1 to 7 months ahead is provided in Fig. 8 using several methodologies such as CNN-RF, ARIMA, MLP, and LR.

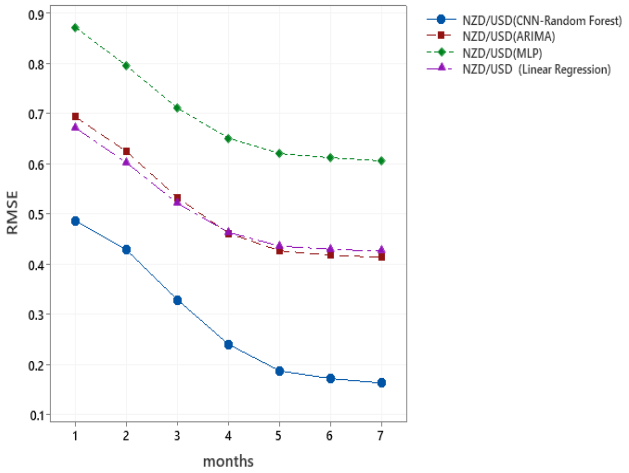


Fig. 8. ER of NZD / USD Seven Months ahead Prediction Error (RMSE).

C. GBP / JPY

Table VII represents, the R^2performance of proposed method with the state-of-the-art techniques is shown. In GBP/JPY ER forecasting, for the first month prediction our suggested technique result is 21.75% bigger than ARIMA, 67.66%bigger than MLP,35.48%bigger than LR. For the second month the result of R^2 of the proposed model is 21.17% bigger than ARIMA, 59.22% bigger than MLP, 35.12% bigger than LR. For the third month the result of R^2 of the suggested model is 20.39% bigger than ARIMA, 65.11% bigger than MLP, 34.40% bigger than LR. For the fourth month the result of R^2 of the proposed model is 19.19% bigger than ARIMA, 75.38% bigger than MLP,35.16% bigger than LR. For the fifth month the result of R^2 of the suggested model is 27.86% bigger than ARIMA,50.05% bigger than MLP,39.95% bigger than LR. For the sixth month the result of R^2 of the proposed model is 31.73% bigger than ARIMA, 50.95% bigger than MLP,42.56% bigger than LR. For the seventh month the result of R^2 of the suggested model is 50.12% bigger than ARIMA, 51.58% bigger than MLP, 50.28% bigger than LR.

The R^2 outcome of the GBP / JPY from 1 to 7 months ahead is given in Fig. 9 using several methodologies such as CNN-RF, ARIMA, MLP, and LR.

TABLE VII. COMPARISON OF R^2 FOR GBP TO JPY FOREX DATA USING DIFFERENT TECHNIQUES

R^2(GBP/JPY)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.9374	0.7699	0.5591	0.6919
2	0.9367	0.7730	0.5883	0.6932
3	0.9367	0.7780	0.5673	0.6969
4	0.9478	0.7952	0.5404	0.7012
5	0.9553	0.7471	0.4771	0.6826
6	0.9573	0.7267	0.4695	0.6715
7	0.9620	0.6408	0.4658	0.6401

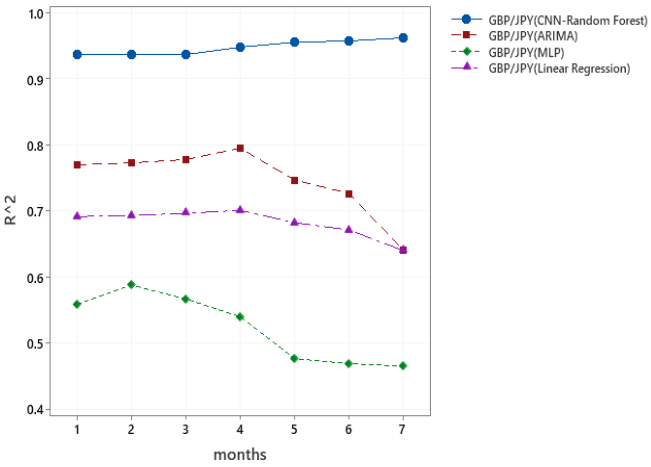


Fig. 9. ER of GBP / JPY Seven Months ahead Prediction Error (R^2).

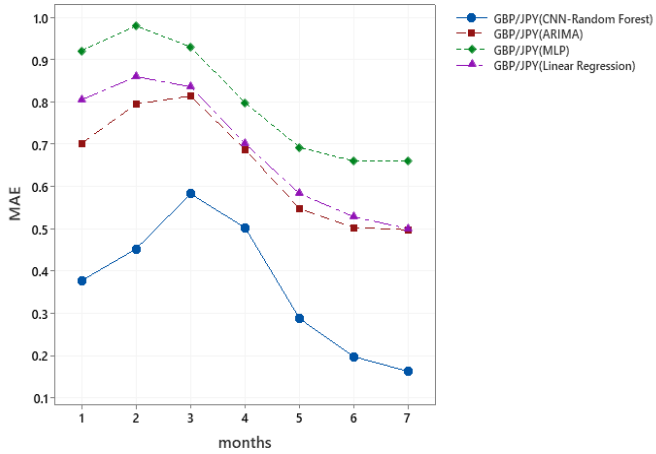


Fig. 10. ER of GBP / JPY Seven Months ahead Prediction Error (MAE).

The MAE (the least value is the better) values for GBP to JPY prediction is indicated in Table VIII. For first month, the MAR value of the proposed model is 46.15% smaller than ARIMA, 58.97% smaller than MLP, 53.04% smaller than LR. For the second month the result of MAE of the suggested model is 43.01% smaller than ARIMA, 53.78% smaller than MLP, 47.31% smaller than LR. For the third month the result of MAE of the proposed model is 28.42% smaller than ARIMA, 37.30% smaller than MLP, 30.27% smaller than LR. For the fourth month the result of MAE of the suggested model is 26.84% smaller than ARIMA, 36.97% smaller than MLP, 28.29% smaller than LR. For the fifth month the result of MAE of the proposed model is 47.43% smaller than ARIMA, 58.41% smaller than MLP, 50.61% smaller than LR. For the sixth month the result of MAE of the suggested model is 60.61% smaller than ARIMA, 69.97% smaller than MLP, 62.48% smaller than LR. For seven month the result of MAE of the proposed model is 67.21% smaller than ARIMA, 75.32% smaller than MLP, 67.37% smaller than LR.

Using multiple techniques such as CNN-RF, ARIMA, MLP, and LR, the MAE outcome of the GBP/JPY from 1 to 7 months ahead is shown in Fig. 10.

TABLE VIII. COMPARISON OF MAE FOR GBP TO JPY FOREX DATA USING VARIOUS TECHNIQUES

MAE(GBP/JPY)				
MONTHS	CNN- RF	ARIMA	MLP	LR
1	0.3784	0.7028	0.9224	0.8058
2	0.4534	0.7956	0.9810	0.8606
3	0.5834	0.8151	0.9305	0.8367
4	0.5034	0.6881	0.7987	0.7020
5	0.2884	0.5487	0.6935	0.5840
6	0.1984	0.5037	0.6608	0.5289
7	0.1634	0.4984	0.6622	0.5009

The seven month Forex prediction for GBP to JPY is mentioned in Table IX. When RMSE(the smallest value is better) performance measured in GBP/JPY ER forecasting for the first month our proposed model result is 33.61% smaller than ARIMA,44.10% smaller than MLP,36.89% smaller than LR. For the second month the result of RMSE of the suggested model is 34.31% smaller than ARIMA,44.69% smaller than MLP,36.75% smaller than LR. For the third month the result of RMSE of the proposed model is 41.40% smaller than ARIMA,52.06% smaller than MLP,43.49% smaller than LR. For the fourth month the result of RMSE of the suggested model is 51.48% smaller than ARIMA,61.74% smaller than MLP,53.13% smaller than LR. For the fifth month the result of RMSE of the suggested model is 59.38% smaller than ARIMA,68.71% smaller than MLP,60.30% smaller than LR. For the sixth month the result of RMSE of the proposed model is 61.73% smaller than ARIMA,70.69% smaller than MLP,62.31% smaller than LR. For seven month the result of RMSE of the suggested model is 63.30% smaller than ARIMA, 71.97% smaller than MLP,63.47% smaller than LR.ARIMA, MLP, LR.

The RMSE outcome of the GBP/JPY from 1 to 7 months ahead is given in Fig. 11 using multiple methodologies such as CNN-RF, ARIMA, MLP, and LR.

TABLE IX. COMPARISON OF RMSE FOR GBP TO JPY FOREX DATA USING DIFFERENT TECHNIQUES

RMSE(GBP/JPY)				
MONTHS	CNN-RF	ARIMA	MLP	LR
1	0.5120	0.7713	0.9160	0.8114
2	0.4648	0.7076	0.8404	0.7349
3	0.3659	0.6245	0.7633	0.6476
4	0.2734	0.5635	0.7146	0.5834
5	0.2178	0.5363	0.6961	0.5487
6	0.2030	0.5305	0.6926	0.5387
7	0.1942	0.5292	0.6930	0.5317

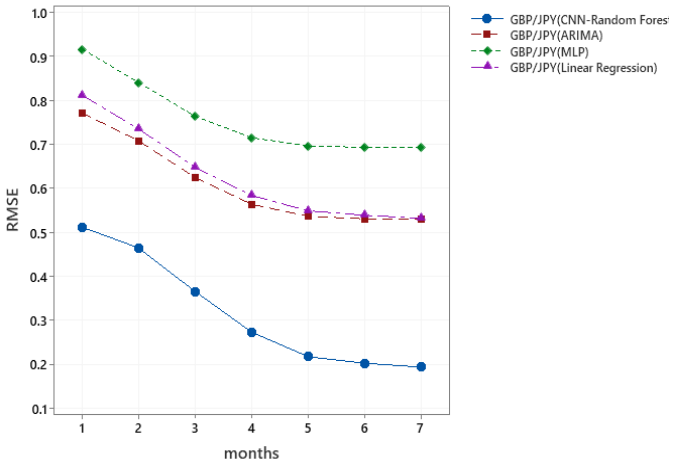


Fig. 11. ER of GBP / JPY Seven Months ahead Prediction Error (RMSE).

D. Performance Comparison

The proposed CNN with random forest model was compared against MLP, ARIMA, and LR layer to discover how good a model is. The forecast for 1 month, 2 months, 3 months, 4 months, 5 months, 6 months and 7 months is created as a proof-of-concept using data from January 2, 2001 to May 31, 2020 for AUD/JPY and GBP/JPY and data from January 1, 2003 to May 31, 2020 for NZD/USD. We have used past every minute opening price, closing price, low price, high price and volume for forecasting the closing price in advance. Tables I to III shows the comparison for 1 to 7 months' ahead prediction in terms of R^2, MAE, RMSE values for AUD/JPY currency pair using different algorithms. The R^2 metric, sometimes called the coefficient of determination, implies that the model is more fit. R^2 can have a value between 0 and 1, with 0 indicating that the model does not fit the given data and 1 indicating that the model fits the dataset completely. The average R^2 value of the proposed model for AUD/JPY is 0.9616. Tables IV to VI compares the R^2, MAE, RMSE values for the NZD/USD currency across a 1 to 7-month time frame. When we examine the results, we can find that the proposed model yields lower MSE, RMSE and MAE over the 1 to 7-month timeframe, implying that the NZD/USD currency pair is more accurate. For the NZD/USD currency pair, the proposed model's average R^2 value is 0.9640. Tables VII to IX compare R^2, MAE, RMSE values for the GBP/JPY currency across a 1 to 7-month time frame. The average R^2 value for GBP/JPY of the proposed model is 0.9620. Hence as all the average value comes near to 1, hence our model is best suitable for all the datasets.

VI. Conclusion with Future Enhancement

Time series forecasting accuracy is crucial to many decision-makers. In this research, the capabilities of CNN model with random forest technique for currency exchange rate forecasting have been investigated and compared with other forecasting methods. After performing an experiments of the exchange rates between the datasets such as AUD/JPY, NZD/USD and GBP/JPY for 1 month, 2 months, 3months,4 months, 5 months, 6 months and 7 months in advance. The model is compared with ARIMA, MLP, LR which clearly establish that the model not only predict the close price but also able to guide the investor to invest in Forex market. Four evaluation criteria such as R^2, MAE and RMSE consider for estimating the performance of the models. Based on the forecasting result our suggested model performs superior than all other models. Furthermore, there are other future research directions for this study. This model will be applied to all remaining significant currency pairs in the future, and the correctness of our suggested model will be estimated. In addition developing a framework for predicting performance based on dynamic data sets which will enhance the exchange rate prediction more efficiently.

References

[1] S.R.Das, D.Mishra, M.Rout" A hybridized ELM-Jaya forecasting model for currency exchange prediction", Journal of King Saud University – Computer and Information Sciences, vol.32, pp.345-366, September 2017.

[2] S.K.Chandar,M.Sumathi,S.N.Sivanandam," Foreign Exchange Rate Forecasting using Levenberg-Marquardt Learning Algorithm", Indian Journal of Science and Technology,vol.9, pp.1-5,February 2016.

[3] T.N.Pandey, A .K.Jagdev, S.Dehuri, S.Cho ,"A novel committee machine and reviews of neural network and statistical models for currency Exchange Rate prediction: An experimental analysis", Journal of King Saud University–Computer and Information Sciences, vol.32,pp.987-999,November 2020.

[4] Z.Chen." The impact of trade and financial expansion on volatility of real exchange rate", Plos one, vol.17, 13 pages, January 2022.

[5] J Sanghun, A. Elsharkawy, and M. S. Kim," Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition", Sensors, vol.22,20 pages,2022.

[6] J.P.Sahoo,A.J.Prakash,P. Pławiak,S. Samantray," Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network", Sensors,vol.22,14 pages, January 2022.

[7] M.Gour,S.Jain,"Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification" ,Computers in biology and medicine, vol.140,16 pages, January 2022.

[8] Y.Ensafi, S.H.Amin, G. Zhang,B.Shah," Time-series forecasting of seasonal items sales using machine learning–A comparative analysis", International Journal of Information Management Data Insights,vol.2,16 pages, April 2022.

[9] E.Hoseinzade, S. Haratizadeh,"CNNPred:CNN-based stock market prediction using several data sources", arXiv preprint arXiv:1810.08923, 39 pages, October 2018.

[10] M.Susruth," Financial Forecasting: An Empirical Study on Box –Jenkins Methodology with reference to the Indian Stock Market", Pacific Business Review International, vol.10, pp.115-123, August 2017.

[11] D.Barbara, C.Li, Y.Jing, A.Samuel," Modeling and Forecast of Ghana's GDP Using ARIMA-GARCH Model", Open Access Library Journal, vol.9, pp.1-16, January2022.

[12] M.M.Panda, S.N.Panda, P.K.Pattnaik,"Exchange Rate Prediction using ANN & Deep Learning Methodologies: A Systematic Review", Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN 2020), India, pp.86-90, February 2020.

[13] P. Escudero, W. Alcocer,J. Paredes," Recurrent Neural Networks and ARIMA Models for Euro/Dollar Exchange Rate Forecasting", Applied Sciences,vol.11,12 pages, January 2021.

[14] M.S. Islam, E. Hossain," Foreign exchange currency rate prediction using a GRU-LSTM Hybrid Network",Soft Computing Letter,17 pages, October 2020.

[15] A.F. Adekoya, I.K . Nti.,B.A. Weyori," Long Short-Term Memory Network for Predicting Exchange Rate of the Ghanaian Cedi", FinTech,vol.1.no.1, pp.25-43,March 2022.

[16] A.A. Baffour,J.Feng,E.K.Taylor," A hybrid artificial neural network-GJR modeling approach to forecasting currency exchange rate volatility", Neurocomputing,vol.365, pp.285-301,August 2019.

[17] P.K. Sarangi, M.Chawla, P.Ghosh, S.Singh, P.K.Singh," FOREX trend analysis using machine learning techniques: INR vs USD currency

exchange rate using ANN-GA hybrid approach", Materials Today: Proceedings,vol.49, pp.3170-3176,January 2022.

[18] U.Chowdhury,S.Chakravarty,T.Hossain,"Short-Term Financial Time Series Forecasting Integrating Principal Component Analysis and Independent Component Analysis with Support Vector Regression", Journal of Computer and Communications,vol.6, pp. 51-67,March 2018.

[19] P. Yaohao, P.H.M. Albuquerque," Non-linear interactions and exchange rate prediction: Empirical evidence using support vector regression ", Applied Mathematical Finance,vol.26, pp.69-100,January 2019.

[20] A.R.Nagpur," Prediction of Multi-Currency Exchange Rates Using Deep Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol.8, pp. 2278-3075, April 2019.

[21] J. Zheng, X. Fu,G.Zheng," Research on exchange rate forecasting based on deep belief network", Neural Computing and Applications,vol.3, pp.573-582,January 2019.

[22] P.F.Pai,C.S.Lin," A hybrid ARIMA and support vector machines model in stock price forecasting", Omega,vol.33,pp.497-505,December 2005.

[23] F. Mansour,M.C. Yuksel,M.F.Akay," Predicting Exchange Rate by Using Time Series Multilayer Perceptron",' 3rd International Mediterranean Science and Engineering Congress,IMSEC 2019', 5 pages,2018.

[24] V.Kumar," Prediction of Foreign Exchange Rate Using Data Mining Ensemble Method", Research Project. National College of Ireland, 2016.

[25] A.C.Bayas," Currency Risk Management: Predicting the EUR/ USD Exchange Rate", Major Qualif. Proj. Years, April 2018.

[26] M.V.Valueva, N.Nagomov,P.Lyakhov, G.V. Valuev," Application of the residue number system to reduce hardware costs of the Convolutional neural network implementation", Mathematics and Computers in Simulation (MATCOM) , vol. 177, pp. 232-243,December 2020.

New Blockchain Protocol for Partial Confidentiality and Transparency (PPCT)

Salima TRICHNI, Mohammed BOUGRINE, Fouzia OMARY
Faculty of Sciences Mohammed V University in Rabat, Department of Computer Science, Rabat, Morocco

***Abstract*—Running behind new technologies is increasingly becoming a non-circumventable requirement for organisms' survival. This is not only a strategy to gain a competitive advantage in the market but it is a determinant key for their continuity and persistence. The Blockchain is at the heart of this technological revolution for which transparency, accessibility to the public and the sense of sharing are fundamental properties of its design. Despite its importance, leveraging this technology in an ethical and secure manner by ensuring confidentiality and privacy is a top concern. Through this work, we try to design a new approach to validate transactions within the Blockchain. Entitled "Protocol for Partial Confidentiality & Transparency PPCT", this new protocol makes possible to seek a compromise between the two requirements: Confidentiality & Transparency. It allows introducing a new notion of confidentiality that we have named partial confidentiality. Subsequently, it applies it on the transactions exchanged while ensuring the process of their validations by the different nodes of the Blockchain. In addition, and through the use of hashing and digital signature functions, this protocol also ensures integrity and authentication within its validation process. To present this work, we will first discuss the state of the art on the different current privacy approaches and our motivation behind this work. Then we will explain more about the different stages of this process, its benefits and areas for improvement.**

***Keywords*—Blockchain; security; privacy; confidentiality; transparency; integrity; authentication; validation process**

I. INTRODUCTION

Blockchain, the founding technologies of cryptocurrency is very fashionable and on trend recently [1]. Its first use took place in 2009 by Satoschi Nakamoto thus giving birth to Bitcoin [2]. And since then, its fields of application have not stopped expanding to serve different sectors, including banks, insurance, the pharmaceutical industry, supply chains. It is at the heart of the current digital technological revolution and considered by some to be the revolutionary successor to the Internet [3]. Indeed, it allows disintermediation or the renunciation of a trusted third party, thanks to its decentralized architecture coupled with its transparency and its high security. The decentralized architecture of the Blockchain results from its constitution as a distributed P2P (peer-to-peer) network. The latter is made up of a set of nodes through which the exchanges and storage of information present in chained blocks (called a chain of blocks) and linked to each other are carried out. The resulting chain of blocks is incremented as soon as new transactions are validated by a set of network nodes according to a precise consensus algorithm (proof of work or proof of stack or...). With regard to security, it is guaranteed fundamentally in the blockchain by cryptographic processes and in particular asymmetric cryptography [4].

For example, in the cryptocurrency field, if Alice wants to send money to Bob (Alice and Bob are the usual protagonists in the cryptographic context) she will create a transaction specifying the amount to be sent and broadcast it to all nodes. This transaction will be grouped with other transactions in a single block, which will be validated later. All of the nodes in this network verify the transactions in this block using a consensus protocol to obtain network approval. As soon as a group of nodes succeeds in verifying and validating all the transactions contained in the block, this later can be added to the Blockchain. Fig. 1 provides an illustration of this process. Thus, when the "block" containing the transaction is approved by the other nodes and added to the Blockchain that this transfer of money between Alice and Bob will become legitimate [4].

Nevertheless, the adoption of this technology encounters some difficulties and obstacles that prevents the putting in practice and still cause concern among entrepreneurs and investors. Serious problems but which cannot be evaded because this innovation offers much needed potential and assets. These issues can be projected on three essential scales, on the one hand the safety of the technical tools and their ability to guarantee the different properties promised by this technology, on the other hand, the functional aspect linked to the business domains of its application, including financial and economic issues. Then, and finally the legal scope reflecting the reliability claimed by the technology to protect public order, control the consumer and eliminate fraudulent use. As part of this contribution, we focus more on the aspect of transaction confidentiality which is one of the major challenges for the adoption of this protocol. Our goal is to seek a trade-off between transparency and data privacy which is a dilemma of Blockchain adoption [3]. Indeed, although the Blockchain is a transparent and public register, keeping transactions or certain sensitive information confidential is one of the great expectations in this technological context [5], which hinders its adoption in most cases. So, there are many legitimate reasons for conducting private transactions. Reasons may be critical, such as revealing your sources of income to your competitors, your health problems, etc. Other reasons are not necessarily critical, such as keeping a surprise for your spouse secret. In any case, it is a human right that we must absolutely respect. Unfortunately, this property is not supported in today's most popular Blockchains such as Bitcoin. In general, in this type of system, pseudonym tools are often used in order to hide the real identity of the users [6].

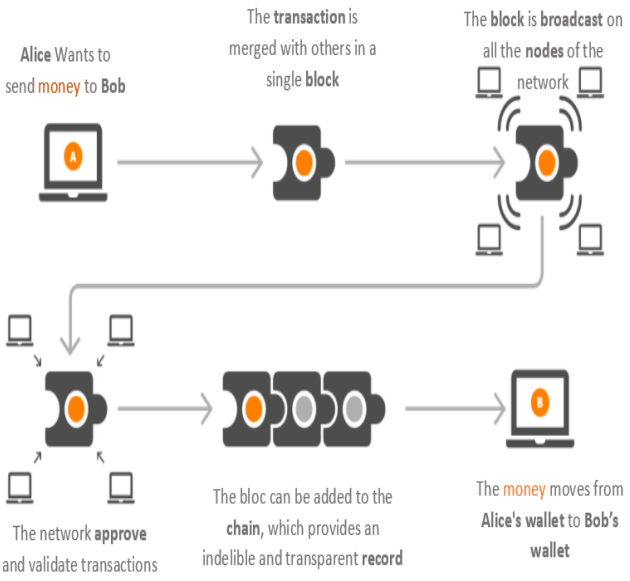


Fig. 1. The Concept of Blockchain Operation.

These tools are based on sender and recipient address type information and solve this problem by relying on the principle of sharing information without being visible to the public. On the other hand, if an adversary has key information about one of the two parties, he can acquire links or relations leading to the true identity and so decrypt the pseudonym of this user [7]. Although the techniques of anonymization and pseudonym are more and more complex and sophisticated; the risks of privacy leakage due to various inference attacks [8] are also well developed [9]. Therefore, it is essential to implement stronger protection mechanisms.

We will dedicate the following section to discuss key techniques that can help improving data confidentiality and transaction privacy on the Blockchain.

II. Related Work

Most of the work that aims to solve the problem of privacy in the Blockchain, is more focused on what is called by "secure computing". Indeed, this type of solution is based on techniques allowing the realization of calculations on data without revealing the secret part of each transaction. To do this, these works are essentially based on one or two of the following encryption approaches:

- The Attribute-Based Encryption (ABE) approach.
- The Secure Multipart Calculation (SMPC) approach.
- The Homomorphic Encryption (HE) approach.
- The Non-interactive zero-knowledge proof (NIZK).

In this section, we will try to address these approaches by describing some of the work done under each approach.

A. Attribute-Based Encryption (ABE)

Proposed in 2005, the concept of attribute-based encryption (ABE) was first introduced based on a single authority [7]. It represents a new encryption policy that builds on the same principle of asymmetric encryption by adding an additional layer to include user-specific attributes. This is a recent and promising approach to provide both data privacy and access control to that data through the integration of private attributes into all the tools of this protocol. The policy for the use of these attributes must be defined in advance by the appropriate authority [10] [11] in order to be able to cryptographically combine decryption keys with data access permissions. The data thus encrypted does not need to be transmitted over a secure channel or stored in a trusted server. To decrypt encrypted data, users must now meet the access policy that is defined based on the attributes that can be associated with data users, data elements, and the environment. Despite their power, these algorithms are not very widespread because of the difficulty associated with their design and implementation. Little is done using this type of encryption [5].

In blockchain, for example, the first proposal for the use of ABE was published in 2011 in [12], it is a decentralized ABE scheme that is based on access tokens assigned according to the rights of each node. Token tracking automatically goes through the processes that are in place. The distribution of tokens no longer relies on a central authority [4], several nodes can be elected through witnesses to play this role. Another ABE-based encryption proposal was patented in [13] and published in 2018. It consists of an encryption solution based on a pre-calculation phase that does not require exchanges with trusted servers and that significantly reduces the cost of computing encryption on devices with limited resources. This encryption is based on CP-ABE "Ciphertext-Policy Attribute-Based Encryption" and defines an access policy for encryption by referring to an access structure in the form of a tree. Encryption goes through 2 steps, the first performs a multiplication between random elements defined by a randomly generated polynomial on each root of the tree and the elements of a cyclic group. The second step is based on the results of the pre-calculation performed previously and stored in a memory to accomplish the encryption task.

B. Secure Multi-party Computing

The Multiparty Computing Model (MPC) is a generic cryptographic scheme for performing secure calculations between two or more parties without revealing their private data inputs. The first variants of this type of encryption were proposed by Andrew Yao [9] [14], the first in 1982 concerns just two parts while the second in 1986 and is generalized on several parts. Other designs of this scheme have been successfully realized and applied on a variety of issues such as distributed voting, private auctions and lately in the Blockchain. In [15], Andrychowicz and all built an MPC protocol for Bitcoin to be used in the lottery field to ensure honest behavior within this Blockchain. This type of algorithm was also used in a work published in [16]. Ce Last offers a secure computing solution for Blockchain networks, it uses the MPC calculation protocol by separating the ownership of the data and the use of this data and allows to reduce the burdens of the computational work to a few nodes by using a layer 2 solution, then it uses the message authentication code (MAC) to verify the accuracy of the calculation carried out. We thus conclude with another work in this same context; this is the

Enigma platform that is also based on SMPC and hardware privacy technology TEE (Trusted Execution Environment) to provide computation over encrypted data and guarantee confidentiality [17].

C. Homomorphic Encryption (HE)

Homomorphic encryption (HE) is a new family of cryptographic tools. It adds a verifiable compute layer while maintaining the confidentiality of source data [4]. Indeed, homomorphic encryption must be able to evaluate encrypted data by performing certain arbitrary functions directly on the ciphertext [18]. On the other hand, when deciphering the results found we end up with values identical to those performed by the same operations on the plaintext. The application of this type of encryption within the Blockchain is of great use to ensure the confidentiality of data. It makes it possible to store the encrypted data in the distributed ledgers of the Blockchain [19], and then execute the validation process on this encrypted data without proceeding to its decryption. Y.Wang and A.Kogan proposed in [20], a new design of a transaction processing system based on the Blockchain and dedicated to accounting and auditing. This design aims to ensure the confidentiality of transactions by using on the one hand the Homomorphic algorithms and on the other hand the approach of the non-interactive proofs with zero knowledge NIZK and more precisely its variant zk-SNARK.

D. Non-Interactive Zero-Knowledge Evidence (NIZK)

Proposed in the early 1980s [21], the ZK zero-knowledge interactive proof system is the first version of this approach that allows a certifier to prove to a verifier that a statement is accurate without providing any useful information to the verifier, in other words, it allows, from a formal proof applied to a secret entry, generate an exit open to the public without disclosure of any other information [22]. This variant then became Non-Interactive in the sense that it no longer requires direct interaction between the certifier and the verifier. It is enough that the latter two share a common reference chain to achieve the same objective, which is zero knowledge. This is called NIZK. The use of this type of algorithm in the Blockchain is in great demand. It has been used in several cryptocurrencies in order to prove the validity of the transfer of the currency between the different entities without having any knowledge about the balance of each entity. Several other versions of this same protocol have been proposed in the context of the Blockchain, the best known of which are currently:

- Zcash [23] a cryptocurrency based on the Bitcoin code and integrates zk-SNARK [24] in order to be able to verify transactions while keeping user information confidential.
- Zerocash over Ethereum (ZoE), applied on Ethereum, it allows a user to store Ether (ETH) in a discreet manner by adding a "serial number" as a commitment in a Merkel tree, which is maintained by the contract [4].

Admittedly, all the approaches just mentioned offer very innovative solutions for ensuring the confidentiality of sensitive data, however, as we can see from the description of each one, these approaches nevertheless remain limited to specific cases and cannot be applicable on any type of consensus [6]. They are more applicable in cases where the process of validating a transaction requires the calculation of one or more operations [16]. So, in the case of a consensus based on a procedural and purely functional smart contract, the application of this type of solutions will not be possible otherwise it will be expensive [17].

Other disadvantages can hinder the use of this type of algorithm such as their slowness and cost in terms of consumption of physical resources [16]. This is due to the fact that the calculations/checks they use are not done directly on the raw data. On the other hand, these algorithms consider a limited amount of data and cannot support a large input volumetric [20]. And finally, these approaches still suffer from their lack of maturity and complexities related mainly to the difficulty of their implementation and implementation [13].

The encryption approach proposed in this work is generic and applicable on any type of consensus. In contrast to the solutions presented above, our model offers more fluidity and makes it possible to exploit the symmetric encryption algorithms that have largely proven their robustness and performance in the field. In addition, it integrates other security tools such as hash functions and digital signatures [25]. The flexibility of this protocol does not only concern the security tools put in place, but also at the level of the distribution of roles at each transaction, something that prevents the vulnerability of the system.

III. Methodology

The new PPCT protocol proposes to keep some of the information confidential and leave other non-sensitive information transparent and readable by everyone. It aims to use the principle of partial confidentiality in order to be able to share transactions in a public way while keeping sensitive data secret except in the eyes of authorized nodes. This solution promotes parallelization of the transaction verification activity to ensure the security of the system by separating the tasks between the different participants of the distributed network.

To further explain the system, below are all the definitions and steps put in place under this protocol.

A. Definition: Partial confidentiality

This work presents a new concept of confidentiality which is partial confidentiality. As its name suggests, this concept aims to ensure the ownership of confidentiality just on a part of the data deemed to be sensitive. To do this, the partial confidentiality algorithm requires prior identification of sensitive information, then it proceeds to ordinary encryption processes to encrypt this data and reintegrate it at the end of the algorithm into its original context, Fig. 2.

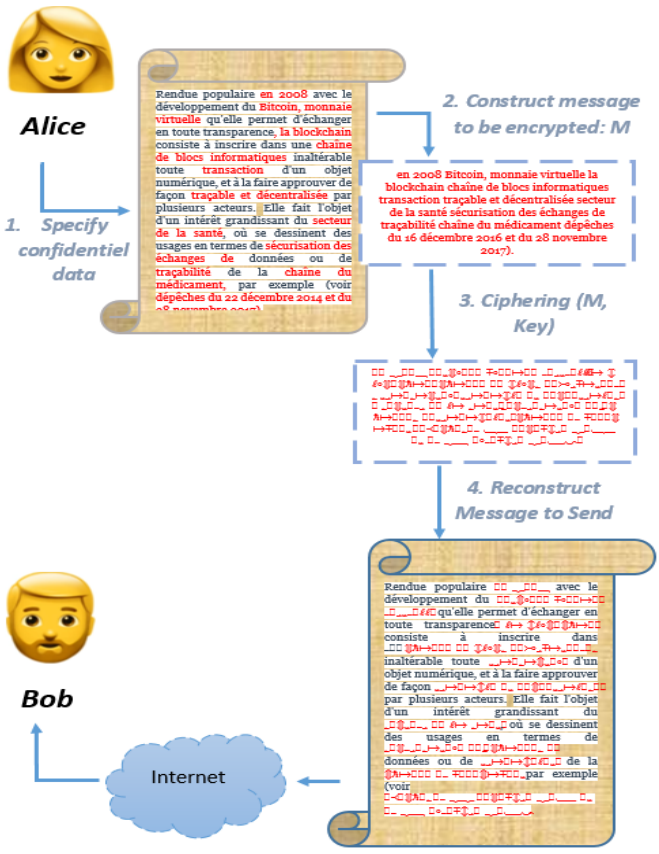


Fig. 2. Principle of Partial Confidentiality.

To better assimilate this algorithm, below are the details of the operations performed at the level of each step:

- Step 1 – Identification: The first step of this system is to identify the sensitive information in the text to be shared using a separation mechanism before and after each part.
- Step 2 - Extraction: The second step is to extract sensitive information from the text to be shared by referring to the separator used for this protocol. Then, they are grouped in an apartment block called black-blocks (BBs) and which will be encrypted in the next step.
- Step 3 – Encryption: Encryption is performed on the black blocks (BBs) containing the sensitive data using a symmetric encryption algorithm whose secret key is that of the entities authorized to validate the transaction or part of the smart-contract.
- Step 4 – Reconstruction: The reconstruction step consists of integrating the different bytes of the encrypted BBs into their initial positions of the clear text by referring another time to the separator set up during the identification step.

B. Transaction Trust Group: TTG

Each participant /organization shall identify in advance its trusted group with which it must ensure the validity of the information exchanged in full transparency. This trusted group can be considered as a private subnet of our distributed network whose cryptographic key exchange is previously carried out outside the Blockchain.

C. Transaction Base

In a Blockchain several types of transactions can circulate and exchange between the different nodes of the network, each transaction reflects a specific functionality in the process to be digitized. It can present a purchase, a transfer, a contract verification, the result of a diagnosis, or others. Regardless of the type of transaction, in our system we require to specify the structure of all possible transactions and give them a well-defined base. The Pillar of each transaction must distinguish between public information and information that must remain secret.

D. Partial Confidentiality

As we have presented before, partial confidentiality is a technique that has just been defined to ensure the privacy of a set of information included in a data model intended for the public. To do this, it is necessary to go through the separation of these two categories of data, encrypt the sensitive part based on a cryptographic encryption tool and finally reintegrate this part into the model in question, Fig. 3.

E. Private Validation

In order to validate the current transaction, each node in the trusted group partially decrypts the model, calculates the hash of the content in clear and then proceeds to the realization of their tasks necessary for their validation.

Validators send their signed and encrypted responses using the hashed of clear text as the encryption key, it is also called the "public validation key" (kv).

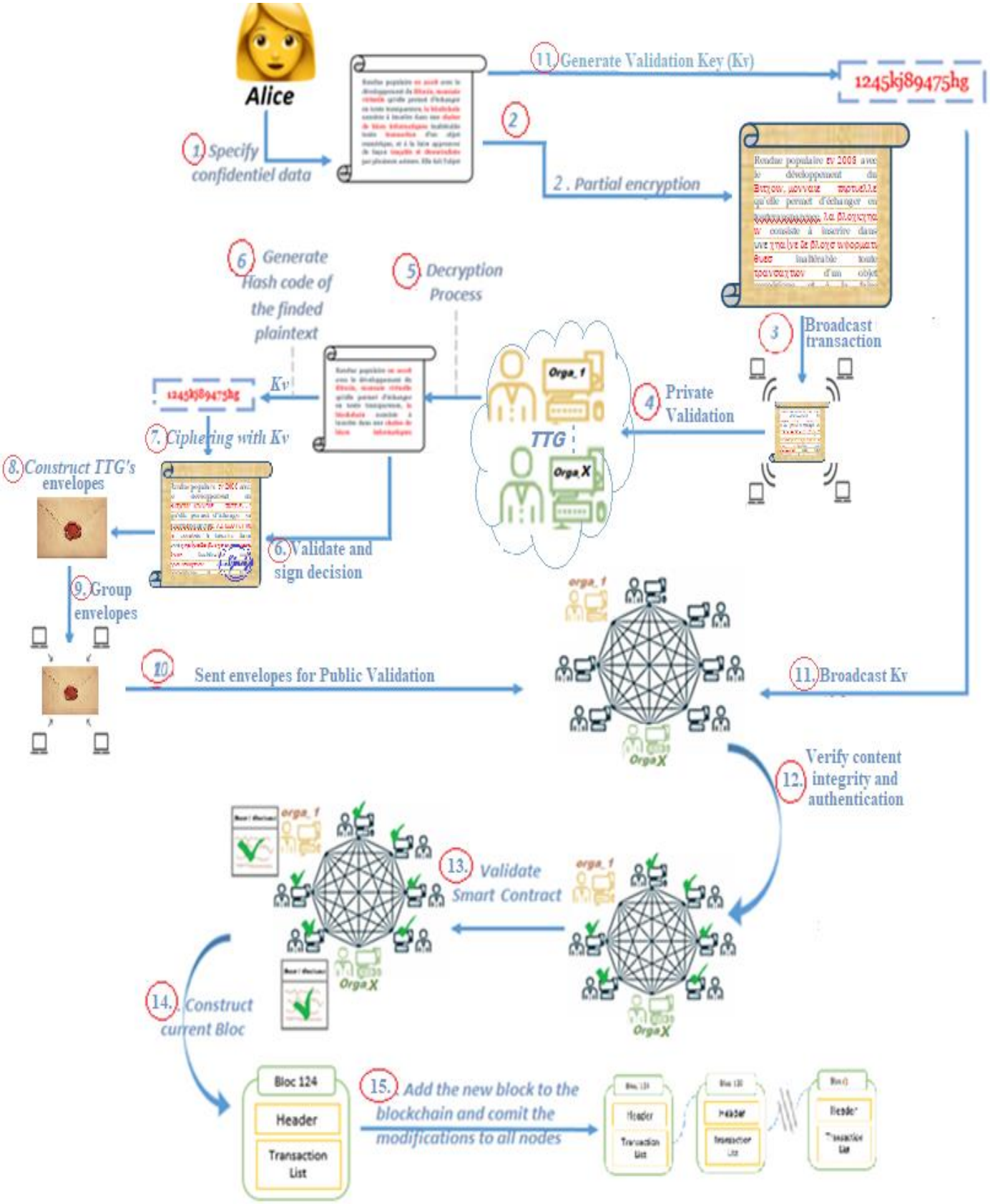


Fig. 3. Protocol Preserving Confidentiality and Transparency (PPCT).

It should be noted that each validator node also has a base for its response and uses its own signature to sign its decision. The whole thing is encrypted using the hash of the clear. This encryption operation has a very important role in this security protocol. On the one hand, it ensures the credibility of the validator's decision because he had the right hash code and therefore was able to rely on the right information contained in the plaintext model. On the other hand, through this encryption we manage to protect the group against the infiltration of malicious nodes so as not to impact the final decision of the transaction, and therefore, we ensure the neutrality of this decision. This operation is considered a closed envelope intended to be broadcast on the network.

F. Public Validation

Once you receive the envelopes from the various validators of the trusted group, a notification is sent to the initiator of the transaction in order to send the hashed clear in the Blockchain. The final transaction is built based on the model, hash and envelopes of the trusted group.

The transaction is added to the Blockchain block and sent to the public network for final validation. The public network opens the envelopes of each validator by performing the decryption operation with the key the hashed of the clear, then compares the results and performs its global checks to validate the operation.

If the public validators manage to decipher the message with the hashed of the clear it means that the private validator had the right content of the model that the initiator of the transaction and therefore on the basis of the different answers the public validates the final transaction.

Once the final validation consensus is passed, the block is stored and added to the ledger via the ordinary mechanisms of the Blockchain.

IV. USE CASE: REIMBURSEMENT FOR CARE

For example, in the field of health where trust holds a dominant and decisive place, the application of the Blockchain can open up very promising horizons and prospects. The desire for perfect traceability in an environment where information is shared and stored in a fairly transparent manner, while ensuring a reasonable degree of security for a real control of the data circulating there. This field of application, which requires the coexistence of these two seemingly antagonistic qualities, is very favorable for the production of our solution. Understanding this health process in a Blockchain is of great use. In fact, it reduces their complexity by facilitating the management of relationships between different heterogeneous information systems while ensuring a secure and transparent exchange of information. This Blockchain brings together the entire medical profession, patients, insurance companies, radiology centers, laboratories, pharmacies, physiotherapy centers.... In short, all professionals in the field of health must come together around this Blockchain, each from its own angle, to ensure the smooth running of different standards put in place in the service of health. These contributions in terms of: trust, security, simplification, and parallelization of verification procedures are immediately and not only in an economic gain in time and money but also in a huge improvement in the quality of life of the populations.

We can imagine a simple and very recurrent use case in our daily life and it is the procedure of reimbursement of care by health insurance. Automating the reimbursement of care is a rather complex task because it depends on the credibility and commitment of all entities in the field of health. The normal course of this reimbursement process requires a complete medical record containing all the supporting documents on the consultations, diagnoses, care and treatments carried out. This proof must be signed and validated by the various interlocutors in the field. All these papers are then deposited with the insurer who in turn carries out its checks and validates the file for reimbursement according to the insured's health insurance contract. Communication between the different interlocutors of a medical file is carried out by the patient. We start from this same principle and we propose the scenario modeled in the sequence diagram in Fig. 4.

Patient

- Identifies its interlocutors in the blockchain.
- Completes their medical file.
- Request for the validation of his file.

Healthcare Professional

- Validation of the medical file by the interlocutors.

Insurance:

- Validation of the medical record by the insurance.
- Reimbursement of the file.
- Validation of the refund transaction by the network.
- Transaction storage.

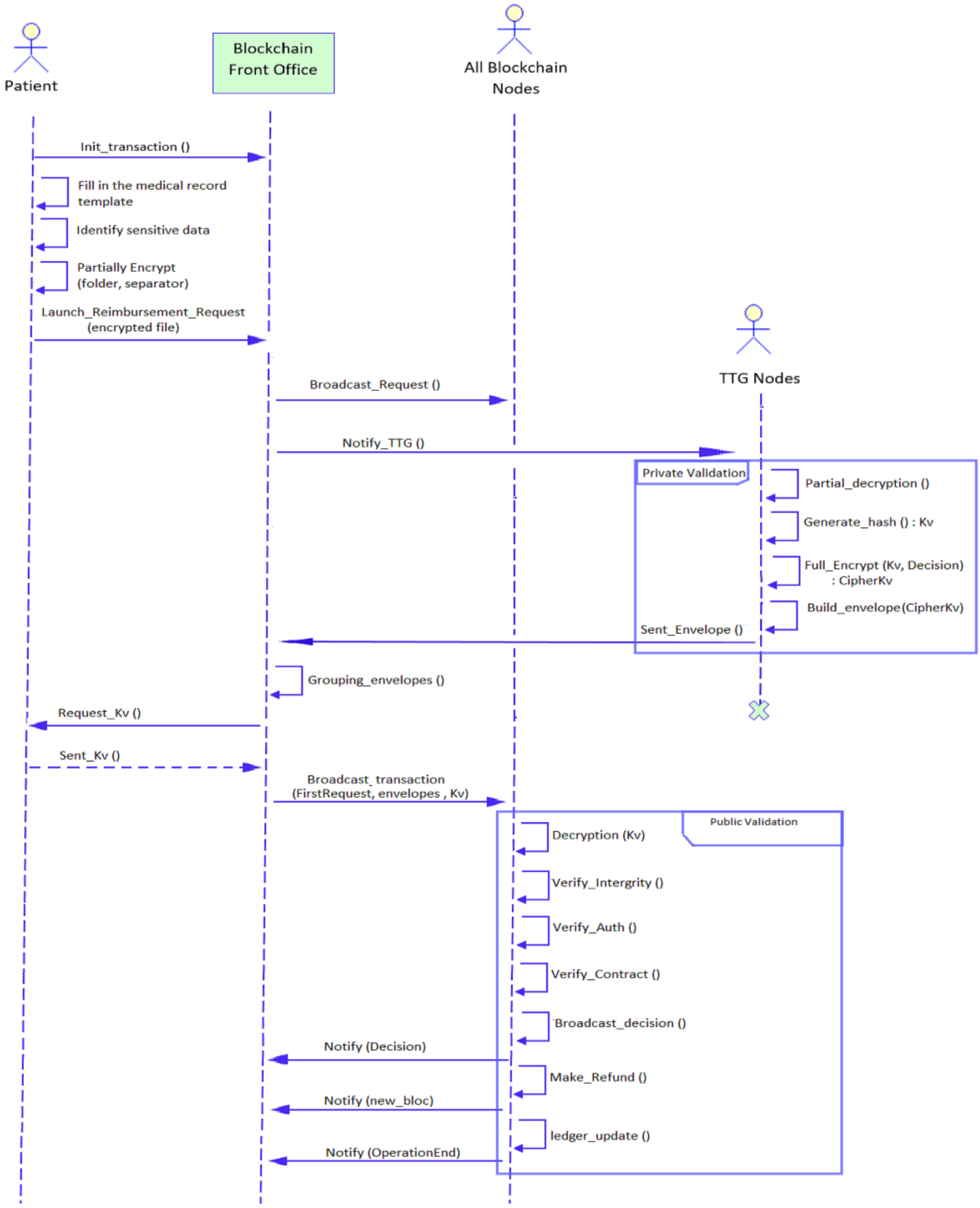


Fig. 4. Sequence Diagram for Healthcare Reimbursement.

V. Result and Discussion

As Blockchain technology is based on a distributed network, the notion of a central authority no longer exists. As a result, all members of this network are invited to execute the consensus in question in order to validate the transactions of this network. Nevertheless, with the immense evolution of the size of current Blockchains, collective intervention becomes more expensive whether in terms of quantity of calculations or transaction fees [26]. Thus, it causes real latency and performance issues on the network [27]. In addition, the security aspect is also impacted and is becoming increasingly difficult to ensure. Our protocol responds to this problem and offers advantages to keep the potential of the blockchain to the maximum while promoting its scalability and security. These benefits can be discussed in the following points:

A. Confidentiality

Confidentiality is ensured based on the process of partial confidentiality using ciphering algorithm [28] . This process makes it possible to encrypt sensitive data not necessary for the public validation of the transaction and then reintegrate it into the overall base of the transaction containing other public clauses. In this way, the public clauses as well as the basis of this type of transaction will be validated in public. On the other hand the trusted group of this transaction guarantee the validation of the private elements as a whole by performing the partial decryption and thus verifying the overall content of the current transaction.

B. Integrity

In addition to the confidentiality of sensitive data, the PPCT protocol also ensures the integrity of this data based on the hash of the clear part [25] [29]. The latter is designated in our protocol by the Kv validation key. This key is the same for all participants in the TTG group. It is used by each member of this group to encrypt all the elements of its envelope and distribute it within the distributed network. Once all envelopes are retrieved, the nodes in the public network ask the initiator to reveal the Kv validation key by distributing it to the network, and then they verify the integrity of the committed transaction. They decipher different envelopes of the TTG group. If the decryption of each envelope passes well it means that the hashed of the clear was the right one for this participant and its validation can be considered in the public validation.

C. Authentication

Ensuring authentication is also one of the advantages of this solution. Indeed, adding the signature [25] [29] of the validator of the TTG group makes it possible to verify his identity and to ensure that the decision sent in the network concerns the right person designated in the TTG group.

D. Off-chain Processing on a Small, Private and Dynamic (Cyclical) Network

The PPCT makes it possible to move part of the work outside the public network. It allows you to create a verifiable property in the outsourced calculation task [6]. This property is ensured by the Kv (Validation Key) which is carried out according to the sensitive data of each transaction. The Kv Key is an effective way for the public to confirm their validation without resorting to sensitive data from that transaction. The network delegates this task to the TTG while maintaining control over the work of this small private network. This solution is very useful in that it applies the principle of off-chain calculation [30] [16] in order to be able to solve this problem effectively. So, he

- does not depend on the size of the blockchain [16];
- promotes the scalability of the blockchain [27];
- reduces calculation costs [16];
- reduces overall transaction validation time [16].

E. Independent Treatment of Public Network Trust

The PPCT does not manage access to the Blockchain in a strict and permanent way, it is only at the time of the creation of the transaction that the initiator of the transaction designates his trusted group. This increases the security of the protocol for two reasons, the TTG group represents entities that are concretely trusted in reality and not defined in a procedural way. The second related to the dynamism of this group so even if one of the malicious nodes manages to integrate the TTG group of the current transaction, it will not necessarily be able to be in the other time. And so, the attack can only concern a single transaction and for which it can be well identified in the public validation stage.

- Permissions are not fixed and change from one transaction to another by favoring the choice of the initiator of the transaction.
- The initiator of the transaction designates his trusted group.

F. Efficient and Secure Processing Capacity

For the PPCT protocol, the calculation and validation time passes faster because it is realized after the decryption of sensitive information by the trusted group. Thanks to decryption [31], validation calculations do not represent any complexity and do not require any requirements in terms of computing power.

G. Independence of Data Type and Size

Digital or literal, Log file or raw data, the PPCT can be applied easily on the different types of transaction unlike other confidentiality protocols that are restricted to numerical data to perform calculations or to some private data to represent identity [32]. In our case, just choose the right Symmetric Encryption algorithms that perform better on the volume and types of data to be exchanged to improve the performance of the protocol [33].

H. Flexibility of Cryptographic Tools

Cryptographic tools can be appropriated according to the constraints of the entities that come into play in this Blockchain [33].

The following table (Table I) summarizes the advantages and disadvantages that we have been able to identify, in order to conclude a static comparison with our work.

TABLE I. COMPARATIVE SUMMARY

Technique	*Advantage*	*Disadvantage*	*Application*
HE	It can perform confidentiality-preserving calculations by performing calculations directly on the ciphertext.	Only certain types of operations, such as addition and multiplication, can be implemented effectively. The computational efficiency of complex functions is very low.	Etherium
MPC	It allows multiple parties to perform calculations jointly on their private data entries without violating their input confidentiality.	Only certain simple functions can be supported, and complex functions are less efficient.	Enigma
ABE	It can simultaneously ensure data confidentiality and precise access control.	Issuing and revoking the attribute certificate in a distributed environment has yet to be resolved.	SO
NIZK	The user can easily prove that he has a sufficient balance for the transfer with NIZK, without revealing the account balance.	Less effective	Zcash
PPCT	Ensures partial confidentiality while maintaining transparency on transaction.et also integrity and authentication	Latency to improve	SO

VI. CONCLUSION

The work presented in this paper has opened a new compromise track between transparency and privacy within the blockchain. First, through the introduction and after giving a general overview on blockchain technology, we have exposed the problem studied during this work as well as the main objective of our proposal which is: the confidentiality of sensitive data while preserving transparency within the blockchain infrastructure. We then presented the various related works in this context that are generally designed at the basis of secure calculation algorithms namely: ABE, SMPC, Homomorphic encryption and NIZK algorithms. Then we gave the detailed description of our solution entitled "Protocol for Partial Confidentiality & Transparency (PPCT)" by discussing, at the end of the article, its advantages over other equivalent systems. In summary, and through this work, it can be concluded that the PPCT protocol proposed here, was able to solve the problem of privacy by exploiting the conventional tools of cryptography and adapting them to the concepts of the blockchain in a new perspective. We defined a new notion of confidentiality that is partial confidentiality and then we integrated hash functions and signature algorithms at the heart of the validation process. The integration of the latter ensures the increase in the level of security of the system by guaranteeing the other two pillars of security, namely: integrity and authentication. Thus, the PPCT protocol makes it possible to ensure the confidentiality of the sensitive data of each transaction via the partial encryption of it, then the integrity of this data using its hash, then the authentication of the first validators' decision-makers of this transaction via their signatures. As presented in the discussion section, the performance of this protocol is well in line with other privacy solutions that rely on secure calculation tools. In our next work, we want to focus more on the latency time between the initial validation and the final validation of the transaction while studying ways to improve this criterion by adding the aspect of parallelism in this process. We also want to carry out comparative experiments between the different possible combinations and proposing a clearer scheme of use on each type of need.

REFERENCES

[1] R. a. C. M.-B. Beck, "Blockchain as Radical Innovation: A Framework for Engaging with Distributed," in Proceedings of the 50th Hawaii International Conference on System Sciences, 5390-5399, 2015.

[2] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," https://bitcoin.org/bitcoin.pdf, 2009.

[3] L. F. Frédéric Lau, Blockchain : passer de la théorie à la pratique. Les enjeux de la transformation pour l'adoption, Cigref, octobre 2018.

[4] I. Bashir, Mastering Blockchain-deeper insights into decentralization, cryptography, Bitcoin and popular Blockchain frameworks, Packt, 2017.

[5] "Rui Zhang and Rui Xue and Ling Liu, Security and Privacy on Blockchain,2019,zhang2019security,arXiv:1903.07602".

[6] P. J. T. C. X. L. a. Q. Xiaoqi Li, "A Survey on the security of blockchain systems," Future Generation Computer Systems, 2017.

[7] "Amit Sahai and Brent Waters. [n. d.]. Fuzzy Identity-Based Encryption. 457–473".

[8] H. C. Y. Z. M. H. M. S. Z. C. Yourong Chen, "A survey on blockchain systems: Attacks, defenses, and privacy preservation," High-Confidence Computing , vol. 2, no. 100048, 2022.

[9] "A. C. Yao. [n. d.]. Protocols for secure computations. In SFCS 1982. 160–164.".

[10] "Sanjam Garg, Craig Gentry, Shai Halevi, Amit Sahai, and BrentWaters. [n. d.]. Attribute-Based Encryption for Circuits from Multilinear Maps. 479–499.".

[11] "Sergey Gorbunov, Vinod Vaikuntanathan, and Hoeteck Wee. [n. d.]. Attribute-based Encryption for Circuits. In Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing (STOC 2013). 545–554.".

[12] "Allison Lewko and Brent Waters. [n. d.]. Decentralizing Attribute-based Encryption. In EUROCRYPT 2011. 568–588.".

[13] "OUALHA, Nouha et Janneteau, Christophe, Méthode De Chiffrement Basée Sur Les Attributs Comprenant Une Phase De Pré-Calcul , Brevet Européan "EP3371929B1"".

[14] "Andrew Chi-Chih Yao. [n. d.]. How to Generate and Exchange Secrets. In SFCS 1986. 162–167".

[15] "Marcin Andrychowicz, Stefan Dziembowski, Daniel Malinowski, and Lukasz Mazurek. [n. d.]. Secure Multiparty Computations on Bitcoin. In SP 2014. 443–458.".

[16] "Zhang, Derek & Su, Alex & Xu, Felix & Chen, Jiang. (2018). ARPA Whitepaper".

[17] O. N. a. A. P. Guy Zyskind, "Decentralized Computation Platform with Guaranteed Privacy.," Computer Science., 2015.

[18] "Craig Gentry. [n. d.]. Fully Homomorphic Encryption Using Ideal Lattices. In STOC 2009. 169–178".

[19] "Marten van Dijk, Craig Gentry, Shai Halevi, and Vinod Vaikuntanathan. [n. d.]. Fully Homomorphic Encryption over the Integers. In EUROCRYPT 2010. 24–43".

[20] "Yunsen Wang, Alexander Kogan,Designing confidentiality-preserving Blockchain-based transaction processing systems,International Journal of Accounting Information Systems,Volume 30,2018,Pages 1-18,ISSN 1467-0895,".

[21] "S Goldwasser, SMicali, and C Rackoff. [n. d.]. The Knowledge Complexity of Interactive Proof-systems. In STOC 1985. 291–304.".

[22] "Manuel Blum, Paul Feldman, and SilvioMicali. [n. d.]. Non-interactive Zero-knowledge and Its Applications. In STOC 1988. 103–112.".

[23] "Eli Ben Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. [n.d.]. Zerocash: Decentralized Anonymous Payments from Bitcoin. In SP 2014. 459–474.".

[24] "Jens Groth. [n. d.]. Short Pairing-Based Non-interactive Zero-Knowledge Arguments. 321–340".

[25] F. G. E. N. S, "Techniques Of Cryptography," CNAM, 2002.

[26] H. L. J. H. X. W. Q. Miao, "An intelligent and privacy-enhanced data sharing strategy for blockchain-empowered Internet of Things," Digital Communications and Networks, no. doi: https://doi.org/10.1016/j.dcan.2021.12.007, 2022.

[27] A. A.-A. H. K. Yehia Ibrahim Alzoubi, "Blockchain technology as a Fog computing security and privacy solution: An overview," Computer Communications, vol. 182, no. https://doi.org/10.1016/j.comcom. 2021.11.005, p. 129–152, 2022.

[28] S. F. M.Bougrine, "Improving Performance of the Symmetrical Evolutionary Ciphering System SEC," InternationalJournal of High Performance Systems Architecture, vol. 10, no. 1, pp. 12-19, 2021.

[29] O. P. V. E. V. S. Menezes A.J., Handbook Of Applied Cryptography, Crc Press, 1997.

[30] U. Kumar, "Understanding Ethereum — Pertinent problems,Scalability, and Possible Solutions.," Coinmonks, 01 06 2018. [Online]. Available: https://medium.com/coinmonks/understanding-ethereum-pertinent-problems-scalability-and-possible-solutions-eb4fec0405be.

[31] S. a. O. F. a. I. A. a. B. M. a. A. M. Trichni, "New intelligent strategy for encryption decisional support system," International Journal of High Performance Systems Architecture, vol. 9, no. 4, pp. 173-181, 2020.

[32] R. P. M. Patel, "Improved Identity Based Encryption System (IIBES): A Mechanism for Eliminating the Key-Escrow Problem," Emerging Science Journal, vol. 5, no. 1, 2021.

[33] F. O. a. M. B. Salima TRICHNI, "New Smart Encryption Approach based on Multidimensional Analysis Tools," New Smart Encryption Approach based on Multidimensional Analysis Tools, vol. 12, no. 5, 2021.

Image-based Automatic Counting of *Bacillus cereus* Colonies using Smartphone

Phongsatorn Taithong[1], Siriwan Wichai[2], Rattapoom Waranusast[3], Panomkhawn Riyamongkol[4]
Faculty of Engineering, Naresuan University, Phitsanulok City, 65000, Thailand[1, 3, 4]
Faculty of Medical Science, Naresuan University, Phitsanulok City, 65000, Thailand[2]

***Abstract*—Substantial amounts of *Bacillus cereus* bacteria present in food indicates that the food is unsafe to eat, so counting B. cereus colonies in food samples is a common test for food cleanliness. Manual counting of B. cereus bacteria colonies requires approximately 2-5 minutes per Petri dish, depending on the number of colonies present. This study presents a new smartphone-based method called Bacillus Cereus Image Counting System (BCICS, "B. kiks") for automatic counting of B. cereus colonies. BCICS uses image processing techniques including Projection Profiling, Circle Hough Transformation, Adaptive Thresholding, and Power-Law Transformation to achieve high image clarity and then uses the Connected-Component Labeling (CCL) technique to correctly count the colonies, including overlapping colonies. These techniques are built into a convenient Android smartphone application. Results of counting the colonies with BCICS were compared with results of hand counting the same dishes. The accuracy rate of each dish count was calculated, as well as the average dish accuracy across all dishes. BCICS counted total colonies with an accuracy of 90.14%, which is close to that of hand counting accuracy since hand counting itself commonly involves an error rate of 5-10%. Importantly, the application took only 3-5 seconds to count one Petri dish, which is more than 74 times faster than the time required for manual counting.**

Keywords*—*Bacillus cereus* bacteria; colonies; automatic counting; android phone application; image processing

I. Introduction

Bacillus cereus, or B. cereus, is an aerobic spore-forming bacterium that naturally inhabits soils and plants [1,2]. B. cereus is one of the most important foodborne pathogenic bacteria, and it can cause two types of food poisoning. The first, emetic syndrome, is characterized by nausea and vomiting and occurs within a few hours of consumption of contaminated food. The second, diarrheal syndrome, is characterized by abdominal pain and diarrhea [3] and occurs within 3-4 hours of consumption of contaminated food. Besides B. cereus, other members of the Bacillus group of bacteria can cause food poisoning outbreaks through consumption of raw fruits and vegetables, and this is a major concern in food safety [4-6].

B. cereus food poisoning most commonly occurs when prepared foods are left to sit without adequate refrigeration for several hours before serving. B. cereus concentrations reaching approximately 10^6 cells/gram of food will typically provoke abdominal pain and bloody diarrhea. There is an incubation period of 4-16 hours following ingestion, with symptoms lasting for 12-24 hours [4]. In Slovakia, B. cereus colonies in pasteurized milk at 30°C are not allowed to exceed a total dish count of 3×10^4 CFU/ml. When pasteurized milk is incubated for 5 days at 6°C, the total dish count must not exceed 5×10^4 CFU/ml. [7].

The Standard Petri Count (SPC) method is a conventional technique used to examine Petri dishes for microbiological growth that is large enough to be seen with the naked eye or with a magnifying glass. SPC can be used to test milk quality. To do this, a sample of the milk is diluted with distilled water and then spread onto a Petri dish that contains a medium where bacteria present in the food will start to grow. Once a Petri dish is inoculated and incubation begins, the bacteria colonies will usually be apparent and ready for counting within 24-48 hours [8-9]. When only a few colonies of B. cereus are found, this indicates that the milk meets common quality standards. Colony counting normally requires physically seeing and tallying the colonies by hand, and the accuracy of the total count for a dish can vary from one counter to another, because the B. cereus colonies can be numerous and appear in many sizes. An experienced counter can count the colonies in one Petri dish in approximately 2-5 minutes, depending on the number of colonies present in the dish.

The current study introduces a new method of B. cereus colony counting in which an image is taken of the Petri dish to be counted. The image goes through a series of image processing techniques, including Color Transformation, Thresholding, Projection Profile, Circle Hough Transform, Adaptive Thresholding, and CCL, which together count the colonies automatically. These procedures are built into a convenient Android smartphone application called Bacillus Cereus Image Counting System (BCICS, "B. kiks") that substantially reduces the time needed to count B. cereus colonies, compared to hand counting.

Deep learning and machine learning are not suited to this study, because they both require large amounts of data. All educational institutions in the region were under strict Covid-19 protocols when the study was conducted, so 20 cultured Petri dishes were obtainable and no more. Analysis was conducted on 420 images from these dishes. Brightness and contrast were variously increased and decreased (from -15% to +15%) on a portion of the images to test the method rigorously under a variety of conditions. Bacteria colonies were counted using image processing techniques.

The purpose of this study is to reduce the amount of time required to count colonies, which can benefit microbiologists,

nutritionists, and others. The paper is organized as follows. Section I provides an introduction. Section II reviews relevant previous studies. Section III describes materials and methods used. Section IV reports the experimental results. Section V is a discussion of the results. Section VI summarizes the study in a conclusion. Section VII acknowledges various people who contributed support and assistance in the course of this study.

II. Related Work

Previous studies on the counting of bacteria colonies have varied in both the type of bacteria counted and the methods employed. Hogekamp et al. (2020) developed an automated image-based system for counting Escherichia coli DH5-Alpha colonies. When each petri dish image was imported into Hogekamp's program, it was divided into 6 equal parts similar to slicing a pizza. The system randomly selected 1 of these 6 parts. After that, five image processing techniques were used to count the colonies: background removal, contrast increase, median filter (to reduce visual noise), brightness increase, and dark maxima detection of colonies. The resulting number of colonies was multiplied by 6 to make up for the unprocessed image parts. The main problem that caused counting errors was the random selection of image parts, because the parts did not all have similar numbers of colonies [10].

Ferrari, Lombardi, & Signoroni (2017) designed and tested two different machine learning approaches to counting colonies. The first method was based on extraction of a complete set of custom-designed morphometric and radiometric features used within a Support Vector Machine (SVM) solution. e second approach was based on the design and configuration of a Convolutional Neural Network (CNN), a kind of deep learning architecture. Both approaches used the same 28,500 images. The study found that the CNN method worked better and had a 91.5% accuracy rate. Ferrari et al. were able to use machine learning because of the large number of images available [11].

Siragusa et al. (2018) developed a semi-automated, image-based colony counting system named CoCoNut (Colony Counter developed by the Nutech department at the Technical University of Denmark). CoCoNut consists of an ImageJ macro program and a 3D-printed photo light box. V79 and HeLa colonies were tested. Results were compared to other available programs. CoCoNut succeeded in distinguishing between freestanding and overlapping colonies. CoCoNut software was also faster than other available programs. However CoCoNut required the user to adjustment program parameters according to colony size. Depending on users to do this correctly is by nature risky and best avoided when possible. CoCoNut's counting results were sometimes inconsistent because of the parameter input requirement [12].

Correa et al. (2020) developed semi-automated counting of the bacterium Bacillus pumilus and the yeast Meyerozyma guilliermondii, both isolated from diesel oil. Correa's method involved a series of steps. First, the original image is resized to a smaller image to save processing time, and this smaller image is then converted to grayscale. The grayscale image is then converted to a binary (black and white) image. A binary opening filter is applied to filter out most of the optical noise. Next, colonies are identified using the luminosity features, because a colony is more luminous at its core than at its edges. Using that fact, the identification method involves computing the points of local maximum luminosity. Correa's method was compared with regular hand counting, and the results showed accuracy of 94.64% and 93.41%, respectively. Correa's system delivers good accuracy, but it runs on computers, not on phones, so it is not particularly mobile [13].

III. Materials and Methods

A. Materials

Three kinds of software were used in this study. The tinkercad.com website was used to design models for 3D printing. Python 7.2 software was used to test the image processing techniques. Finally, Android Studio 4.2 software was used to write the BCICS phone application. Materials used consisted of a photo light box, a Petri dish box, an Android smartphone (OPPO Reno 2 with 48-megapixel camera), a phone stand, and Petri dishes with B. cereus colonies. The photo light box, the Petri dish box, and the phone stand were designed using tinkercad.com and the objects were then formed using a 3D printer (Flashforge Finder 2.0; Flashforge; Jinhua City, Zhejiang Province, China). Fig. 1 shows an actual representation the photo light box arrangement.

B. Methods

The new BCICS method to count B. cereus colonies in Petri dishes involves three stages: the input stage, the processing stage, and the output stage.

1) Input stage: B. cereus colonies from milk samples were cultured on Plate Count Agar (PCA) in Petri dishes in advance at a microbiology laboratory. All Petri dish images in this study were taken using the previously described photo light box and smartphone camera. The images had a focal length of 4.77 millimeters, an aperture of f/1.7, and an ISO in the range of 700-799.

Sample images showing the color, size, and shape of B. cereus colonies can be seen in Fig. 2. While some of the colonies are quite large, others are small enough to be difficult to see with the naked eye, so a magnifying glass is used for hand counting.

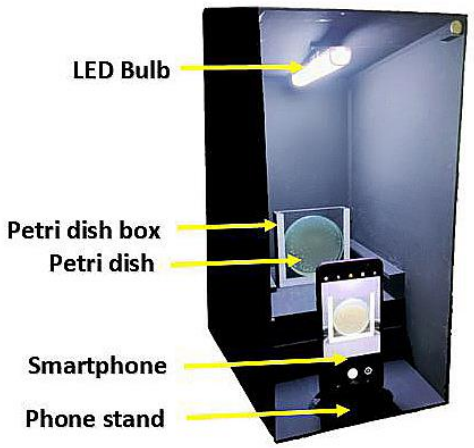
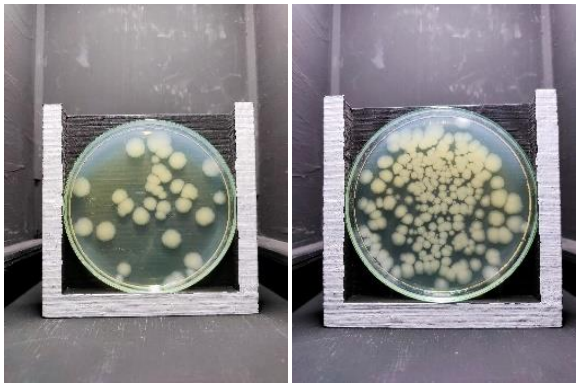


Fig. 1. Actual Representation of the Photo Light Box Arrangement.



(a) (b)

Fig. 2. Sample Images of B. Cereus Colonies in Petri Dishes, showing how the Denseness of the Colonies has been Intentionally Varied, with (a) Sparsely Inoculated and (b) Densely Inoculated.

2) Processing stage: The prepared images are processed in seven steps: isolate the Petri dish box, isolate the Petri dish, remove heightened brightness, reduce turbidity, remove the background, mark the colonies, and count the colonies. These steps are shown in Fig. 3.

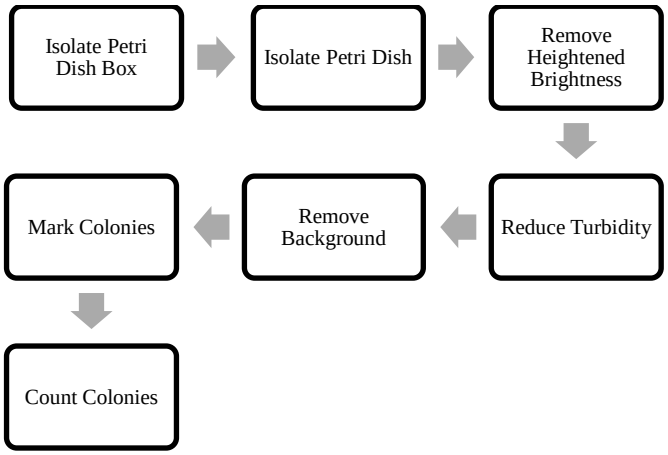


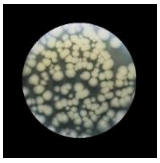
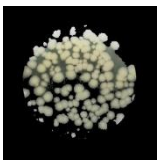
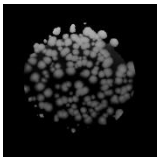
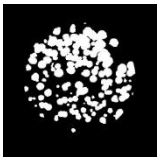
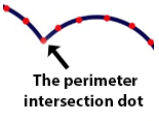
Fig. 3. The Seven Image Processing Steps Carried out by BCICS.

All seven steps involve image processing techniques. The first six steps prepare the image for colony counting, and the last step is where the actual counting occurs. Each of the seven steps in the processing stage is explained in more detail below in Table I.

a) Isolate Petri dish box: Whenever possible, it is good to remove irrelevant parts of an image that could potentially cause errors when looking for the B. cereus colonies. In this first step, the Petri dish box is detected and then everything outside the box is removed. Apart from reducing potential errors, this step also centers the Petri dish in the image. In order to isolate the petri dish box, the image will first be converted to grayscale. The grayscale image will then be converted to a binary image. Finally the Petri dish box can be isolated from the binary image.

Each of the original color images has three color components: red, green, and blue (RGB). Each pixel in the color image is defined by three integers in the range 0-255, which represent the intensity of the three colors red, green, and blue, respectively. The conversion to grayscale is carried using the luminosity method (also called the weighted method) rather than the average brightness method, because the luminosity method considers that the pixel value weight of each color component is different. This solves the problem of the image having three different color components (RGB) that have three different wavelengths. Therefore the luminosity method gives better results than the average method, which tends to produce a darker image. The luminosity method is also the most common option for conversion to grayscale [14-15]. After the image is converted to grayscale, each pixel is instead defined by a single integer in the same range, but with 0 representing black, 255 representing white, and each integer in between representing a corresponding shade of gray.

TABLE I. THE SERIES OF STEPS IN THE PROCESSING STAGE OF BCICS

Step	Action	How It Works	Image Result
1	Isolate Petri Dish Box	White pixels are summed horizontally and vertically to define the four corners of the Petri dish box and enable cropping.	
2	Isolate Petri Dish	The Petri dish is detected from its circular shape and then all the pixels outside the dish are turned to black.	
3	Remove Heightened Brightness	The V component of the HSV color model is used to isolate the band of heightened brightness and then the pixels in the band are turned to black.	
4	Reduce Turbidity	Contrast is increased to overcome the effect of turbidity.	
5	Remove Background	The difference between the colonies and the background is accentuated by converting all pixels to black or white.	
6	Mark Colonies	Red dots are plotted along the perimeter of the colonies.	
7	Count Colonies	Overlapping colonies are identified and all colonies are counted.	

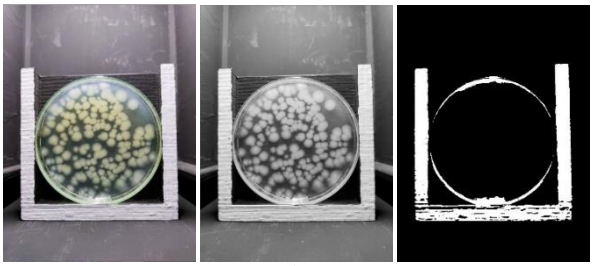
After the grayscale image is ready, the conversion of the grayscale image to a binary image is accomplished using the thresholding technique. This means designating all pixel values as either 0 (black) or 1 (white). When a grayscale pixel value is below a chosen threshold value, that pixel is set to 0. When the grayscale pixel value is equal to or higher than the threshold value, that pixel is set to 1. An example of an image converted from color to grayscale and then to binary is shown in Fig. 4.

The Petri dish box is isolated in the binary image by using the projection profile technique. This means that the number of white pixels in the binary image is summed vertically and horizontally to define the corners of the Petri dish box. These coordinates in the form of pixel positions are then used to remove the irrelevant area outside the Petri dish box. Examples of the horizontal and vertical projection profiles of a Petri dish box image are shown in Fig. 5.

Specifically, the sum of white pixels on the horizontal axis is used to define the top and bottom edges of the Petri dish box, and the sum of white pixels on the vertical axis is used to define the left and right edge of the Petri dish box. On the horizontal axis, the top edge of the Petri dish box can be determined as the last row from the top at which the sum of white pixels is still zero, and the bottom edge of the Petri dish box is the last row from the bottom at which the sum of white pixels is still zero. Similarly, on the vertical axis, the left edge of the Petri dish box is the last column from the left where the sum of white pixels is still zero, and the right edge of the Petri dish box is the last column from the right where the sum of white pixels is still zero. Using these four reference points, the original color image is cropped. The binary image is discarded. A sample result is shown in Fig. 6.

b) Isolate Petri dish: If the Petri dish box is not removed from the image, errors will occur as a result of visual noise caused by the rough surface of the box. The circular shape of the Petri dish makes it suitable for use with the Circle Hough Transform (CHT) technique, which can detect circular objects in an image. The CHT technique accepts a radius range specified by the user and then scans an image, looking for circles with a radius that falls within the designated range [16]. In the current study, the radius range to seek was defined as 150-200 pixels, which corresponds to the radius of a Petri dish. Once the circle of the Petri dish was found by CHT, all pixels outside of the circle were set to 0 (black) in order to isolate the dish image from background, as shown in Fig. 7.

c) Remove heightened brightness: A heightened band of brightness just inside the perimeter of the dish, as seen in Fig. 8(a), can contribute to errors later when the background is removed, because the brightness makes it more difficult to distinguish the bacteria colonies from the background. This band can be removed using the HSV color model, which consists of three components: hue (H), saturation (S), and value (V) [17-18]. The current study, uses only the V component, shown in Fig. 8(b), which represents the degree of brightness of a pixel, ranging from 0 to 255, where 0 (black) is low brightness and 255 (white) is high brightness. The V component is used for thresholding here to create a binary image. Any V component pixel value that is greater than or equal to 60 is set to 1 (white), and values less than 60 are set to 0 (black). An example is shown in Fig. 8(c). The white pixels resulting from thresholding are then set to black in the original image to remove the band of heightened brightness. The Petri dish image after removal of the heightened brightness is shown in Fig. 8(d).



(a) (b) (c)

Fig. 4. A Sample Petri Dish Image Converted from (a) Color to (b) Grayscale and then to (c) Binary.

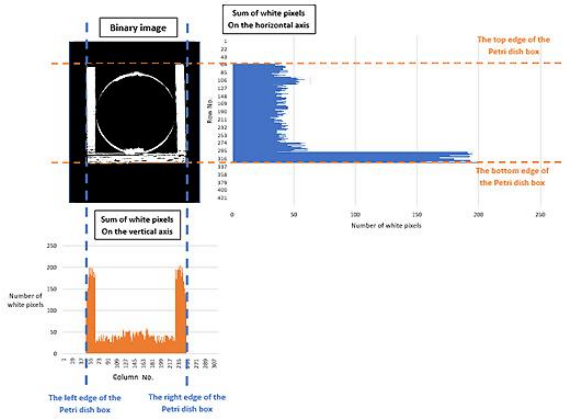


Fig. 5. Example of the Horizontal and Vertical Projection Profile of a Petri Dish Image.

Fig. 6. After using the Projection Profile Technique, the Cropped Image Contains only the Petri Dish and the Petri Dish Box.

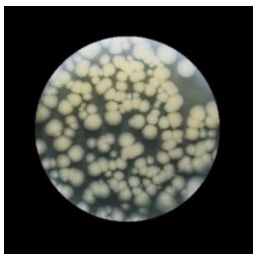


Fig. 7. After use of the CHT Technique, the Petri Dish is alone in the Image, with only Black Pixels outside the Dish Edge.

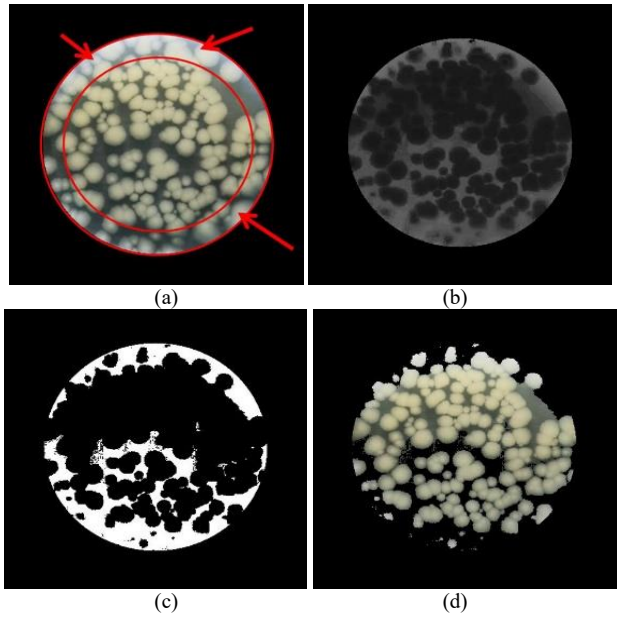


Fig. 8. Removing the Band of Heightened Brightness from the Petri Dish Image: (a) Area of Heightened Brightness, (b) V Component of Previous Image, (c) Binary Image from V Component, and (d) Image after Removal of Heightened Brightness.

d) Reduce turbidity: The culture medium has a natural turbidity from the variety of chemicals it contains, as seen in Fig. 9(a). Turbidity, too, makes removal of the background more difficult later because it decreases contrast in the image. Power-Law transformation is therefore used to overcome the effect of the turbidity by increasing contrast in the image, as shown in Fig. 9(b).

e) Remove background: Removing the background involves setting the background pixels to 0 (black) and the colony pixels to 1 (white). However, while regular thresholding was a useful technique earlier to prepare for the projection profile, the level of visual noise it generates is too high for the current situation, which is more sensitive. Therefore, Adaptive Thresholding (AT) is used to remove the background. Unlike regular thresholding, AT has no set threshold value. The threshold value is dynamic and changes in order to adapt to different brightness environments in different parts of the image [19]. AT works in the following way. Every pixel in the image is processed one by one. At each pixel, a square block of surrounding pixels at which the target pixel is at the center are examined for their brightness. The average brightness of the block is calculated, and the threshold, which is inputted by the user only as a constant, is adjusted automatically to the average brightness of that block before determining if the target pixel will be set to black or white. When the target pixel is at or near the image edge, the missing part of its block that would theoretically extend outside the image is automatically ignored, and the average brightness is calculated from the remaining portion of the block that can still be constructed within the image. In the current study, the block size was set to 555 x 555 pixels, and the AT constant value was set to -25. An example of a Petri dish image after completion of this step is shown in Fig. 10.

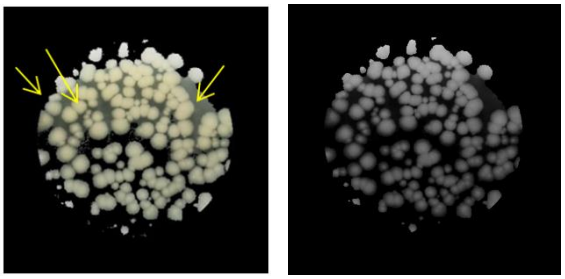


Fig. 9. Petri Dish Image (a) with Natural Turbidity in the Culture Medium and (b) after Turbidity was decreased using the Power-Law Transformation.

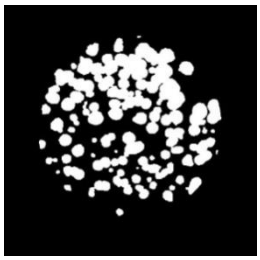


Fig. 10. A Petri Dish Image after use of Adaptive Thresholding to make the Colonies Stand out more clearly from the Background.

f) Mark Colonies: Colonies currently appear in a binary image as clusters (i.e. blobs) of white pixels against a black background. Since the colonies are of many sizes and often overlap, the Connected-Component Labeling (CCL) technique is now used. First CCL finds all the blobs of white pixels [20]. However this is not enough to distinguish between overlapping colonies. So CCL then marks the perimeter of each blob with red dots. Fig. 11 shows these red perimeter dots superimposed on the original petri dish image, with the red dots surrounding each colony or group of overlapping colonies.

Fig. 11. Standalone Colonies and Overlapping Colonies Marked with Red Dots on their Perimeter after using the Connected-Component Labeling Technique.

g) Count Colonies: All of the previous steps have prepared the image for the actual counting of the *Bacillus cereus* colonies, which happens now. When there are overlapping colonies, CCL will always place a red dot at each intersection of the perimeters of the overlapping colonies. Most of the red perimeter dots are not at perimeter intersections, but wherever there is an intersection, CCL will put a dot there. CCL can use these perimeter intersection dots (PIDs) to distinguish and count overlapping colonies with the Convex Hull Detection function. Fig. 12 schematically represents two overlapping colonies, showing sample

placement of red dots along the perimeter and the two PIDs for this configuration.

A total of 1-2 PIDs represents 2 cell colonies, 3 PIDs represent 3 cell colonies, 4 PIDs represent 4 cell colonies, and so on. The number of colonies in the whole Petri dish is summed for the final result.

3) Output stage: The previously described Input Stage and Processing Stage have been incorporated into a convenient and user-friendly tool, the BCICS android application. This is where the results of the first two stages are processed and displayed to the user. The BCICS app was built with Android Studio (Version 4.2.2) and OpenCV library (Version 3.4.14).

When the user launches BCICS, they are prompted to take a petri dish photo. If for any reason the user wishes to retake the photo, they can do so as many times as desired. As soon as the user confirms the photo, the application processes the image and displays the total number of colonies counted. At this last step, the user can choose either to photograph another Petri dish or to exit. A flowchart of the application is shown in Fig. 13.

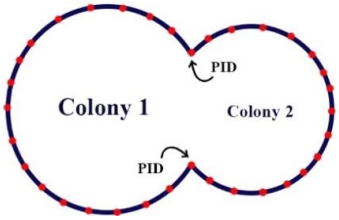


Fig. 12. A Schematic Representation of Two Overlapping Colonies, showing Red Dots Placed along the Perimeter and the Two Perimeter Intersection Dots (PIDs) for this Configuration.

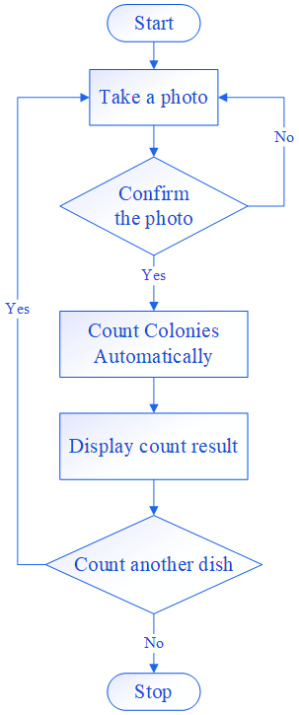


Fig. 13. Flowchart of the BCICS Smartphone Application.

(a) (b) (c) (d)

Fig. 14. Steps in using the BCICS Application: (a) Home Page, (b) Photography Page, (c) Photo Confirmation Page and (d) Result Display Page.

A sample usage of the BCICS app is shown in Fig. 14. The application opens with the home page in Fig. 14(a). Pressing Start takes the user to the dish photography page in Fig. 14(b). Taking a dish photo brings the user to the photo confirmation page in Fig. 14(c). Acceptance of the photo initiates the automatic counting of the colonies, and the user is taken to the result display page in Fig. 14(d), where both the total number of colonies in the dish and the colony locations are displayed.

IV. Results

The BCICS application counted the bacteria colonies in each Petri dish in an average of 4.44 seconds. The minimum time was 3.50 seconds, and the maximum was 5.41 seconds. Hand counting of the same dishes for comparison took an average of 5:32 minutes. The minimum time was 53.04 seconds, and the maximum was 6:51 minutes. Previous studies report that hand counting takes most people 2-15 minutes per dish. The percentage error was calculated according to the formula shown in Equation 1:

$$\frac{hand\ count - BCICS\ count}{hand\ count} \times 100 \quad (1)$$

Fig. 15 shows the distribution of various error rate ranges across all dishes. The error rate range with the most incidences is 0-3.6. The average error rate across all dishes is 9.86% and the standard deviation of the error rate is 7.64. In comparison, the error rate for hand counting falls in the range of 5-10%, because accurate hand counting depends on many factors [21].

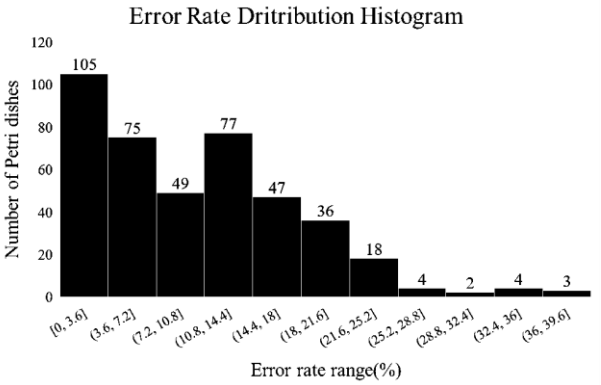


Fig. 15. Histogram Showing a Spectrum of Error Rate Ranges and the Number of Petri Dishes inside each Range.

V. Discussion

Although BCICS delivers high accuracy results, at present some errors are still unavoidable. Within the 8.53% error rate, overcounting represents 66.15% of the errors and undercounting represents 33.85%. Interestingly all errors can be traced back to only two steps: Step 3 (Remove Heightened Brightness) and Step 6 (Count Colonies). All the overcounts occurred in Step 6, while most of the undercounts occurred in Step 6, with some in Step 3 as well. The most common cause of overcounting is the presence of a gap in a colony group. An example of such a gap is shown in Fig. 16. These gaps create inadvertent extra PIDs that disrupt the formula for counting colonies based on the number of PIDs. This issue is a common limitation of CCL. Koomsap et al. (2014) also encountered this issue, and they developed a topological hierarchy-contour tracing algorithm to handle nests of interconnected contours [22]. It might be possible to adapt Koomsap's solution to BCICS as well.

The main cause of undercount errors is another kind of unusual formation of overlapping colonies that interferes with the system of counting PIDs to count colonies. Undercounting can happen when curves of two or more overlapping colonies are so similar and closely aligned to each other that they appear to be a single curve with no perimeter intersection and therefore no PID. This kind of missing PID leads to an undercount error. An example of such a configuration is shown in Fig. 17. Fig. 17(a) shows a common configuration of two overlapping colonies. The perimeters of the two colonies are distinguishable enough that CCL can sense the need for placement of two PIDs, and the colonies are counted correctly. However, Fig. 17(b) shows a less common configuration in which two colonies are even closer, and their perimeters have almost merged. In this unusual case, CCL may not sense the need for PIDs, so this configuration can generate an undercount error. The overlapping of colonies or other objects is a common issue that other researchers have tried to solve in various ways, including the erosion algorithm (a morphological image processing technique) [23], the watershed transform algorithm [24], and the distance transform algorithm [25]. Unfortunately none of these alternative techniques were found to give better results that CCL.

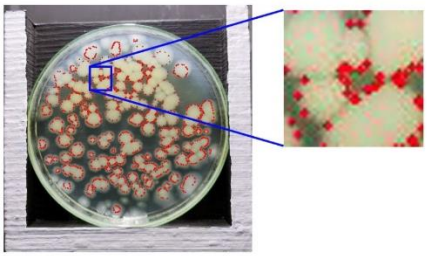


Fig. 16. An Example of a Gap inside a Colony Group, which Causes Creation of Inadvertent Extra PIDs that Disrupt the Formula for Counting Colonies.

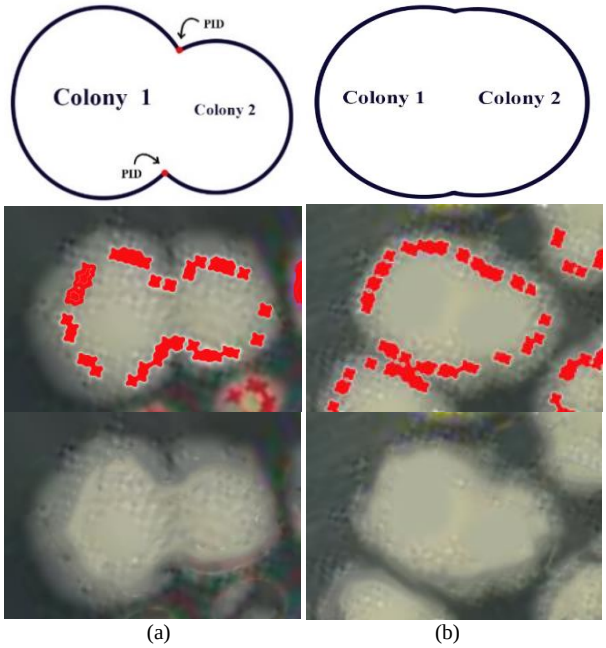


Fig. 17. Examples of Overlapping Colonies in which (a) the Perimeters are Sufficiently Distinguishable to Generate PIDs and be Counted Correctly and (b) the Perimeters are too Similar to Generate PIDs and an Undercount can occur.

A less common cause of undercount errors is unintended removal of a colony from the image during image processing step 3, when the HSV color model and thresholding are used to remove the band of heightened brightness. An example is shown in Fig. 18. Recently, Yifei et al. (2021) encountered an issue similar to this one when processing photos of car license plates. Reflection from another car's headlights created glare and thus areas of excessive brightness in the photos. Yifei applied tone mapping, and this technique was effective in enhancing the local contrast and preserving the image details and structures [26]. It is possible that tone mapping might also be useful in the current study.

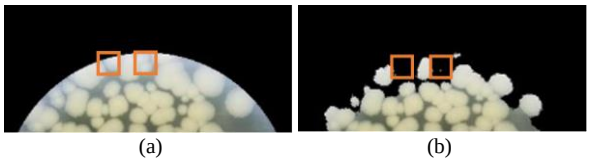


Fig. 18. Examples of Colonies Inadvertently Lost from an Image during Processing Step 3 are shown (a) before Removal of the Heightened Brightness Band and (b) after Removal of the Band.

VI. Conclusion

The BCICS application developed in this study successfully reduces the time needed to count *Bacillus cereus* bacteria colonies in a Petri dish from the average of 5:32 minutes required by hand counting to an average of 6.04 seconds in this automated method. By integrating and automating the seven steps involved, BCICS makes the counting process intuitive and easy to carry out. Because the application utilizes image processing rather than less nimble methods such as densitometric software, it can handle any number of Petri dishes images, even if there are only a few. Although BCICS makes counting B. cereus much faster than hand counting, its accuracy rate of 90.14% is high and comparable to hand counting. In future research, use of the morphological operation called erosion might enable highlighting of inconspicuous perimeter intersections of overlapping colonies and thereby raise the performance of BCICS yet further.

Acknowledgment

The authors are grateful to Ms. Nootsara Yinyom, master's degree student in the Faculty of Medical Science, Naresuan University, for her invaluable assistance in the microbiology laboratory. Thank you also to Mr. Paul Freund of Naresuan University Writing Clinic, Division of International Affairs and Language Development (DIALD) for his help proofreading and editing this manuscript.

References

[1] Sever, B., U.S. Food and Drug Administration (FDA). 2020, Salem Press.

[2] United States, F. and M. Drug Administration. Division of, Bacteriological analytical manual. 1969, Washington.

[3] Can, H.Y., et al., Psychrotrophic properties, toxigenic characteristics, and PFGE profiles of Bacillus cereus isolated from different foods and spices. Propriedades psicrotróficas, características toxigênicas e perfis PFGE de Bacillus cereus isolado de diferentes alimentos e especiarias., 2022. 52(4): p. 1-11.

[4] Antequera-Gómez, M.L., et al., Sporulation is dispensable for the vegetable-associated life cycle of the human pathogen Bacillus cereus. Microbial Biotechnology, 2021. 14(4): p. 1550.

[5] Jovanovic, J., et al., Bacillus cereus food intoxication and toxicoinfection. Comprehensive Reviews in Food Science & Food Safety, 2021. 20(4): p. 3719-3761.

[6] Jessberger, N., et al., *The Bacillus cereus Food Infection as Multifactorial Process.* Toxins, 2020. 12(11).

[7] Valik, L., F. Gorner, and D. Laukova, Growth dynamics of Bacillus cereus and shelf-life of pasteurised milk. 2003, MINISTRY OF AGRICULTURE CZECH REPUBLIC: Czechoslovakia. p. 195-202.

[8] Pisuttilap, N., & Saengswetmaneengam, N., Study of Escherichia coli analysis in frozen seafood by MPN technique / Suksa kan wikhro Escherichia coli nai ahan tha-le chae khaeng duai withi MPN technique. Bulletin of the Department of Medical Sciences (Thailand), 2001. 43(2): p. 95-101.

[9] Puchalt, J.C., et al., Active backlight for automating visual monitoring: An analysis of a lighting control technique for Caenorhabditis elegans cultured on standard Petri plates. PLoS ONE, 2019. 14(4): p. 1-18.

[10] Hogekamp, L., S.H. Hogekamp, and M.R. Stahl, Experimental setup and image processing method for automatic enumeration of bacterial colonies on agar plates. PLoS ONE, 2020. 15(6): p. e0232869.

[11] Ferrari, A., S. Lombardi, and A. Signoroni, Bacterial colony counting with Convolutional Neural Networks in Digital Microbiology Imaging. Pattern Recognition, 2017. 61: p. 629-640.

[12] Siragusa, M., et al., Cell colony counter called CoCoNut. PLoS ONE, 2018. 13(11): p. 1-18.

[13] Correa, C., et al., Use of digital images to count colonies of biodiesel deteriogenic microorganisms. Journal of Microbiological Methods, 2020. 178.

[14] Kanan, C. and G.W. Cottrell, Color-to-Grayscale: Does the Method Matter in Image Recognition? PLoS ONE, 2012. 7(1): p. 1-7.

[15] Wu, T. and A. Toet, Color-to-grayscale conversion through weighted multiresolution channel fusion. Journal of Electronic Imaging, 2014. 23(4).

[16] Gonzalez, R.C. and R.E. Woods, Digital image processing. 4th ed., Global edition ed. 2008: Pearson.

[17] Shamshad, A., Building Computer Vision Applications Using Artificial Neural Networks : With Step-by-Step Examples in OpenCV and TensorFlow with Python. 2020, Berkeley, CA: Apress.

[18] Ding, C., et al., Traffic Image Dehazing Based on HSV Color Space. 2021, IEEE. p. 5442-5447.

[19] Prasad Reddy, P.V.G.D., *Blood vessel extraction in fundus images using hessian eigenvalues and adaptive thresholding.* Evolutionary Intelligence, 2021. 14(2): p. 577-582.

[20] Damayanti, F., Y.K. Suprapto, and E.M. Yuniarno, *Segmentation of Javanese Character in Ancient Manuscript using Connected Component Labeling*. 2020, IEEE. p. 412-417.

[21] Jarvis, B., Statistical aspects of the microbiological examination of foods. 3rd ed. ed. 2016: Elsevier.

[22] Koomsap, P. and N. Chansri, Topological hierarchy-contour tracing algorithm for nests of interconnected contours. International Journal of Advanced Manufacturing Technology, 2014. 70(5-8): p. 1247-1266.

[23] Glagoevs, J. and K. Freivalds, Statistical Method for Object Counting. 2017. p. 61-64.

[24] Arámbula Cosío, F., et al., Automatic analysis of immunocytochemically stained tissue samples. Medical and Biological Engineering and Computing, 2005. 43(5): p. 672-677.

[25] Zhang, J., et al., A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches. Artificial Intelligence Review: An International Science and Engineering Journal, 2021: p. 1-70.

[26] Yifei, W., et al., High-Brightness Image Enhancement Algorithm. Applied Sciences, 2021. 11(11497): p. 11497-11497.

Anomaly-based Network Intrusion Detection using Ensemble Machine Learning Approach

Abhijit Das[1]
Research Scholar, Dept. of CSE
PES Institute of Technology & Management
Affiliated to VTU, Shivamogga, India

Pramod[2]
Associate Professor, Dept. of ISE
PES Institute of Technology & Management
Affiliated to VTU, Shivamogga, India

Sunitha B S[3]
Associate Professor, Dept. of CSE
PESITM, Affiliated to VTU
Shivamogga, India

***Abstract*—In this study, an Intrusion Detection System (IDS) is designed based on Machine Learning classifiers, and its performance is evaluated for the set of attacks entailed in the UNSW- NB15 dataset. UNSW- NB15 dataset contains 2,540,226 realistic network data instances and 49 features. Most research uses a representative sample of this dataset with present training and testing subsets, which includes 257,673 records in total. The dataset was submitted to visual data analysis to discover potential reasons or flaws which likely challenge Machine Learning classifiers. Pre-processing strategies are necessary before this data can be used for data-driven prototype development for IDS because of the class representation imbalance with pattern counts and feature overlap. The method used for pre-processing is implemented by min-max scaling in the normalization phase, followed by applying Elastic Net and Sequential Feature Selection (SFS) algorithms. This work employed ensemble methods using three base classifiers, namely Balanced Bagging, XGBoost, and RF-HDDT, augmented to address the imbalance issue. Parameters of Balanced Bagging and XGBoost are tuned for the imbalanced data, and the Hellinger distance metric supplements random Forest to address the limitations of the default distance metric. Two new algorithms are proposed to address the class overlap issue in the dataset and applied during training. These two algorithms are leveraged to help improve the performance on the testing dataset by affecting the final classification decision made by three base classifiers as part of the ensemble classifier, which employs a majority vote combiner. The performance evaluation of the proposed method for binary and multi-category classification was evaluated using standard metrics, including those generated from the confusion matrix, and compared to other studies using the same dataset. The proposed design outperforms those reported in the literature by a significant margin for binary and multi-category classification cases.**

***Keywords*—Machine learning; ensemble method; intrusion detection system; UNSW-NB15 datasets**

I. Introduction

Cybercrime has increased dramatically due to the rapid development of technology and the broad distributed use of internet networks around the world. If 2019 has taught us anything, it's that no firm, no matter how big or little, is immune to a cyberattack. Cyber-attacks have become more sophisticated, difficult to detect, more targeted than ever before. Security must be constantly upgraded as a result. It is vital to network security to have a network intrusion detection system (NIDS) in place, as it alerts the right authorities when an incursion is detected. It is undeniable that we are becoming increasingly dependent on Internet every passing day due to human creativity and innovation. At the same time, the world of cybercrime has been populated with the criminals looking for a perfect crime such as stealing confidential information, data, funds or causing harm to target computing infrastructure. They explore possible avenues through the cyberspace and formulate different strategies called cyberattacks, to gain unauthorized access to the computing systems. These strategies (aka attacks) may be highlighted as follows: Distributed Denial-of-Service (DDoS), Man-In-The-Middle (MITM), Password-Based Attacks like Brute Force, Dictionary, Shoulder Surfing, Phishing, Malware like Virus, Trojan, Worm, Rootkit, Ransomware, Spyware, Botnet, Key Logger, Adware, SQL Injection, Cross-Site Scripting (XSS), Eavesdropping, Social Engineering [1]. The attackers leverage the attacks to get access to system resources and either destroy them or collect valuable information.

To reduce cyber threats or recover from damages caused by cyberattacks, the organizations assess the cyber risks. Any uncertainties related to the data resources and computer devices that threatens the confidentiality, availability and integrity of the information or information systems are identified as cyber risks [2]. Confidentiality represents the system resources protected from unauthorized access. While integrity preserves any piece of information from unauthorized changes, availability guarantees that the authorized users can always gain access to the required data resources [3]. According to study the most commonly reported cyberattack is Ransomware, it is predicted that a Ransomware attack may affect a business every 11 minutes and the concomitant damage reach $20 billion. The total loss caused by cybercrimes are also projected to rise by $6 trillion by 2022 [4, 5].

With the help of Intrusion detection system, malicious network traffic and device activity standard security system might not be able to see can be found and blocked. IDS is extremely successful in detecting, identifying, and monitoring threats, to put it more exactly. It is very important for keeping computer systems safe from threats that could harm their availability, integrity, or secrecy [6]. Traditional methods use a predefined database of known attacks and signature patterns to evaluate network packets. The user couldn't use the system because of the possibility of a new zero-day that isn't represented by any of the signatures detected in the database [7]. When it comes to detecting zero-day attacks, existing IDS systems have been found to be ineffective [8]. Because of this, it may be concluded that no matter how precise an intrusion detection (ID) technology is, malevolent attempts can degrade IDS stability. The main contributions of this work include:

- The development of a novel ensemble-based classification model for detecting intrusions that uses data from the UNSW-NB15 dataset.
- A cross-comparison of several methods for selecting features has been done.
- The suggested IDS has been tested to see how well it performs for binary and multi-class classification

II. Related Work

An ML classifier's accuracy can be improved by picking features that can represent incursion patterns. Multiple classifiers can reduce false positives and produce more accurate classification results than a single classifier, according to the research [9].

Kumar et al. [10] developed the unified intrusion detection system (UIDS) by generating the new training and test subsets out of UNSW-NB15. They utilized the k-means clustering algorithm to increase the attack sensitivity as the k-means clustering algorithm was able to identify the similarities between different attack classes. In each cluster, the number of records in some type of attack classes was more than the rest. The authors randomly picked 65% of the records of the dominating class categories to form a training dataset. The remaining 35% of the instances were used to build the test set. They also used information gain algorithm for the feature selection phase. 13 features out 47 were selected due to the improvement of accuracy scores by C5, Chi-Squared Automatic Inference Detection (CHAID), Classification and Regression Tree (CART) is also known as Decision Tree (DT) and Quick Unbiased Efficient Statistical Tree (QUEST) algorithms. These algorithms were used to form the proposed UIDS model. Their study reported 77.87% and 79.12% for the average sensitivity and F-measure, respectively. It also offered 3.80% false alarm rate for Normal instances and 86.15% attack sensitivity.

The authors in [11] implemented MLP as an anomaly detection system for binary classification. They employed RFE along with the Random Forest classifier for the purpose of dimensionality reduction. This method selected the top four informative features. The MLP-based IDS scored 85% for sensitivity and 89% for accuracy on the test subset of UNSW-NB15: 15% of the attack traces and 2% of the normal records were misclassified.

Bayu et al. [12] applied Gradient Boosted Machine (GBM) on three datasets including UNSW-NB15. No feature selection technique was implemented. All 47 features were kept for training and testing phases. The results showed that GBM outperformed four other algorithms, namely RF, Deep Neural Network (DNN), SVM and CART, with the average accuracy value of 93.64% and missed alarm rate of 0.0206 where GBM performance was evaluated on NSL-KDD, UNSW-NB15 and GPRS datasets. While running GBM on UNSW-NB15 alone provided 95.08% accuracy and 2.97% false alarm rate using 10-fold cross-validation and the accuracy of 91.31% and false alarm rate of 8.60% using the Hold-Out method on the original UNSW-NB15 train and test subsets.

In [13], nominal features were converted to numerical and then the Min-Max normalization method was utilized to scale down the values to the range of 0 to 1. They used 5-fold cross-validation without resampling to generate the test and training subsets. They calculated the average of the sensitivity, false alarm rate and accuracy of 5 folds. The authors utilized SVM by taking advantage of hyper clique property of hypergraph to improve the performance of SVM. This optimization technique implemented the feature selection as well. The SVM algorithm was trained with the entire 47 features and then the results were compared with the case when SVM was trained with the optimal number of feature subsets. The optimal feature subset is not reported. However, the number of optimal features is in the range of 30 to 35. They concluded that the feature selection had significant influence on the proposed model which delivered 98.47% sensitivity and 2.18% false alarm rate. However, 94.11% accuracy and 2.18% false alarm rate suggests a relatively large value for the missed alarm rate, which is not reported.

ML and DL are used to develop various IDS systems. Due to the enormous dimensionality of the data, improving IDS efficiency and classification prediction requires feature selection from the complete dataset. Chaouki et al. [14] employed genetic algorithm and logistic regression methods to choose the optimal collection of attributes for NIDS. They used KDD99 and UNSW-NB15 datasets for analysis. The primary purpose of their work was to locate the subset of features with the highest classification accuracy and the smallest number of features. RF, C4.5 and NB Tree are used in the classification stage to evaluate the performance of the generated features. Their results show an accuracy of 99.81% for the KDD99 and 81.42% for the UNSW-NB15 dataset. The results conclude that the UNSW-NB15 dataset is more complicated than the KDD99 dataset. In order to enhance the classification accuracy for the UNSW-NB15 IDS datasets, we must thus attempt various methodologies.

Tian et al. [15] addressed the issues of overfitting and inadequate classification accuracy. They proposed a model based on DBN by using probabilistic mass function encoding and the Min-Max normalisation technique. The proposed method achieves 96.17% and 86.49% accuracy on the NSL-KDD and UNSW-NB15 public datasets. They did not explore the class overlap problem of the UNSW-NB15 dataset, and the selection of DBN values was challenging to acquire without experimentation. It was challenging to choose the best parameter impacting detection accuracy.

The efficient classification of network traffic has been hampered by repetitive and unnecessary data attributes and this problem can be solved by feature selection method to recognize the most important elements. M. S. Abirami et al. [16] proposed Least Square Support Vector Machine (LSSVM-IDS) feature selection methods for IDS. With a 95% accuracy rate, this LSSVM system has correctly predicted the output 95% of the time. Ensemble learning was applied to the UNSW-NB15 dataset using a stacking classifier approach. They used logistic regression as a meta-classifier to integrate RF, SVM, and NB algorithms, and they reached a 95% accuracy rate. When selecting features, the author should have utilised an embedded process that might have improved accuracy. There is no mention of the problem of overlapping classes following feature selection. The class overlap problem, embedded methodology, and feature selection can further improve the

proposed model's accuracy.

A dynamically scalable ML-based NIDS was proposed by Soulaiman et al. [17] to solve the imbalanced class problem using SMOTE. The author has not mentioned the class overlap problem in the UNSW-NB15 datasets, decreasing the attack prediction accuracy in real-time.

In [18], the authors proposed the ensemble extreme learning machine (ELM) along with one-vs-all method to generate multi-class classification model. The proposed algorithm is a combination of a single hidden layer feedforward neural network and a softmax layer to make a multi-class prediction out of an ensemble of single output which was 0 or 1. This algorithm scored 95.66% average accuracy. Also, the authors implemented ExtraTree classifier in order to reduce the dimensionality of feature space. Accordingly, 21 features were selected. In the final stage, weighted extreme learning machine (WELM) was implemented and the accuracy of each attack type increased but still need more improvment. The training was done on 80% of the original training set and the remaining 20% was used for validation to avoid overfitting. The entire UNSW-NB15 test subset was utilized for the test phase.

D. Papamartzivanos et al. [19] combined the decision tree and genetic algorithm to generate classification rules and called their model Dendron. Wrapper technique was used for feature selection, which resulted in 23 being selected as informative features. The authors reported the sensitivity of 97.39% for Normal records and the average false alarm rate of 2.61%. In this study, 10% of the instances for each of the 9 classes were considered for building the training set and the remaining 90% was kept for the test set. To address the imbalance problem of multi-class classification in UNSW-NB15, 50% of Worms attack class records were included in the training subset and remaining 50% was kept to test the model.

The integrated rule-based model in [20] is trained to detect five class types to avoid overlapping. These rules were generated from four tree-structured classification algorithms, C5, CHAID, CART and QUEST. Training and test sets were built by eliminating some instances from the original training and testing subsets using k-means clustering. These instances belong to Analysis, Backdoor, Fuzzers, Shellcode and Worms attack types. These attack types suffer from overlapping problem and their presence may cause poor results. For the feature selection phase, the genetic algorithm was used and 22 features were picked accordingly. They reported good accuracy and sensitivity values for the Normal records. However, the average accuracy and sensitivity are 93.94% and 65.21%, respectively.

III. Proposed Methodology

A. Data Preprocessing

Data preprocessing transforms raw data into an appropriate framework or format before processing using a learning algorithm. These techniques significantly impact Machine Learning (ML) and Deep Learning (DL). This section explains these strategies and how they help to maximize the proposed ensemble model's performance.

Data Cleaning: It removes duplicate or irrelevant entries or features from the dataset. The UNSW-NB15 has no duplicate characteristics or records. However, all 49 functionalities are not required to be used. IP address and port number all work together to identify a computer's infrastructure. Other characteristics, such as record start time and record last time, may not have a great deal of use either. Because of overfitting, the Machine Learning (ML) model may not generalize correctly if these attributes are retained. The attack_cat feature was used for multiclass classification, and the label was used for binary class classification.

Data Transformation: Nominal features make up three of the 41 features that aren't targeted. Nominal features have transformed into integers to make ML models and scalers easier. The label encoder is used to convert the nominal features to numerical features. All the features have transformed to the same range of values using scalers, which helps most learning algorithms weigh in the entire features equally. The nearest shrunken centroid has been measured for each attack class. The Euclidean and Mahalanobis distance was used to measure the distance between the attack class centroids. The transformation algorithm which maximizes the intra-cluster distances has been identified and used to address the overlapping problem. According to the finding, the min-max scaler increases the distance of the centroids of four attack classes (Backdoor, DoS, Generic, and Reconnaissance) away from the Normal class centroid more than other methods like Normalization, Robust Scaler, Standardization, Quantile and Power transformation. Consequently, min-max scalar has been employed in this study.

Feature Selection: Fourteen different feature selection algorithms were implemented on the dataset to extract the informative as well as representative features. Chi-squared, Information gain (tree-based feature selection), CFS, ReliefF, and mRMR, are employed among filter-based feature selectors; genetic algorithm, Recursive Feature Elimination (RFE), and Sequential Feature Selection (forward selection and backward elimination) are picked as wrapper methods; and Lasso and Elastic Net are chosen among embedded feature selectors. The effect of feature selection algorithms on the classifier performance was assessed and evaluated using ensemble algorithms including Random Forest, Bagging, Balanced Bagging (BB), AdaBoost, XGBoost, Gradient Tree, Extremely Randomized Trees (ERT), Easy Ensemble (EE) and many other algorithms such as naïve Bayes, SVM and MLP Neural Network using Back Propagation (BP) optimization algorithm. Performance evaluation of classifiers on the dataset, which was preprocessed with the set of feature selection algorithms, indicated that the two best feature selection algorithms are Elastic Net and Sequential Forward Selection (SFS) running in conjunction with Random Forest, Bagging and XGBoost classifiers. The Elastic Net feature selection algorithm combined with the Balanced Bagging machine learning classifier and the SFS feature selector with the Random Forest classifier and XGboost performed the best among all possible combinations tested when considering the F1-score as the metric. In conjunction with the Balanced Bagging classifier, the Elastic Net feature selector performed the best for 24 features from the datasets. In conjunction with the Random Forest classifier, the SFS feature selection algorithm performed the best for 8 features that form a proper subset of 24 features.

TABLE I. MISSED ALARM RATE VALUES FOR THE SET OF 11 CLASSIFIERS

Classes	SVM	NB	Bagging	NN	RF	ERT	AdaBoost	GT	BB	XGBoost	EE
Analysis	1.00	1.00	0.99	1.00	0.86	1.00	0.87	1.00	0.77	0.87	0.86
Backdoor	1.00	1.00	0.93	0.98	0.76	0.95	0.93	0.95	0.59	0.64	0.60
DoS	0.90	0.99	0.88	0.98	0.87	0.87	0.99	0.93	0.81	0.83	0.99
Exploits	0.90	0.97	0.21	0.82	0.21	0.28	0.70	0.09	0.42	0.04	0.99
Fuzzers	0.94	0.63	0.42	0.89	0.39	0.42	0.93	0.55	0.32	0.29	0.75
Generic	0.51	0.03	0.03	0.03	0.03	0.03	0.42	0.03	0.04	0.02	0.42
Normal	0.04	0.65	0.24	0.24	0.21	0.23	0.86	0.34	0.34	0.39	0.70
Recon	0.65	0.71	0.19	1.00	0.17	0.22	0.17	0.21	0.17	0.18	0.99
Shellcode	0.96	0.98	0.31	1.00	0.30	0.52	0.92	0.60	0.06	0.12	0.81
Worms	1.00	0.95	0.86	1.00	0.04	0.86	1.00	0.56	0.09	0.43	0.88

B. Classifier Development Methodology

This research uses the training and testing subsamples existing on the UNSW website. There are 175,341 records in training and 82,332 records in testing data subsets. Each of which contains the records belonging to 9 different attack classes consisting of Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, Worms, and the Normal class. Exploits, Generic and Normal instances make up a large portion of subsamples at 16%, 22% and 38% of the overall records in the training subset, respectively. In extreme cases, Worms attack instances form a mere 0.06% of training and test sets. Since the number of records of majority classes significantly outnumber the records of minority classes, there exists a severe class imbalance problem.The overlapping problem is present in these datasets between the Normal class and some attack classes. This work develops a design to address these problems and an ensemble classifier to identify the threats as normal or attacks.

The model has been implemented by numerous machine learning classifiers such as SVM, naïve Bayes (NB), multilayer perceptron (MLP) neural network (NN), Bagging, Random Forest (RF), Extremely Randomized Trees (ERT), AdaBoost, Gradient Tree (GT), Balanced Bagging (BB), XGBoost, and Easy Ensemble (EE) on the dataset for the case where the classifier algorithms employed all of the original features without implementing data normalization methods. The results are shown in Table I in terms of missed alarm rate (MAR) which is one of the, if not, most critical performance metrics for intrusion detection context. Table I shows that Balanced Bagging, XGBoost and Random Forest lead in their performances for this dataset with respect to the MAR metric. Consequently, these three classifiers will be employed in the design of an ensemble classifier.

The model also employ the Hellinger distance criterion [21] along with the Random Forest classifier to improve the quality of split. Decision Trees are easy to code, interpretable, fast, and nonlinear. However, they suffer from overfitting, axis-parallel splitting and skewness sensitivity. The overfitting problem is mitigated by tree pruning, while axis-parallel splitting can be addressed by building a forest of orthogonal and oblique decision trees. Skewness sensitivity of decision trees arises due to utilizing some popular splitting criteria including information gain and Gini measure. Hellinger distance measure can address this problem due to its skew insensitivity property. For instance, suppose that the model have two classes and Random Forest is applied on the training subset with 175,341 records. Also, in this scenario, the model have 1% of the entire data in class 'A' and the remaining records are in class 'B'. In the case that Random Forest splits the data on the feature 'rate' using a test or threshold value of 200,000, one splitting scenario for such an imbalanced dataset could be as presented in Fig. 1.

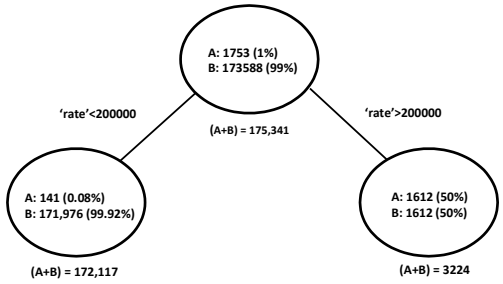


Fig. 1. An Example to Illustrate Hellinger Distance Metric Utility

In fact, regardless of balanced or imbalanced property of a dataset, the best split is made for binary classification when the entire data points in class 'A' are placed in the left node and all data points in class 'B' are placed in the right node. If that is the case, the perfect score for Entropy and Hellinger distance would be 1.0 and $\sqrt{2}$, respectively. In this example, which is a reasonably good but not a perfect split, the scores measured by Entropy and Hellinger are tabulated in table II.

TABLE II. ENTROPY AND HELLINGER DISTANCE SCORE MEASUREMENTS

	Entropy	Hellinger
Perfect Score	1.0	$\sqrt{2} \simeq 1.41$
Evaluated Score	0.015	1.29

According to score, Hellinger distance takes this split into account to form the final prediction. In contrast, given the Entropy formula as

$$E = -\sum_{c=1}^{n} p_c \log_2 p_c \tag{1}$$

and the IG formula as

$$Gain = E(parent) - \frac{W_L}{W_{parent}} E_L - \frac{W_R}{W_{parent}} E_R \tag{2}$$

The split in Fig. 1 does not appear to be "desirable". In equations (1), pi is the probability distribution associated with

class c, and c={1,2,...,n}. Information gain(IG) in equation (2) is the difference of Entropy in parent node (E_{parent}) from the Entropy of its left (E_L) and right (E_R) child where w_L is the number of data points in the left node, w_R is the number of data points in the right node, and wparent is the number of data points in the parent node.

For this split, it looks very probable that a record coming into the right node will be misclassified. This misclassification probability is $\frac{3224}{175341}$(2% of the entire data), while 98% of the data will be most probably classified correctly if they find their way into the left node. Accordingly, it can be considered as a good split. However, the information gained by Entropy measurement indicates that this split is a very bad one since the score is 1.5% of the perfect score shown in Table II. All the while, the Hellinger distance metric values this split by assigning it 91.49% of the perfect or maximum achievable score. Hellinger distance metric thus addressed, for the most part, the problem of skewness sensitivity if utilized by a decision tree classifier and will likely improve the performance of such an algorithm.

C. Training Methodology

The training set is split into two subsets: one subset, aka training subset, is used to train the classifier and the second subset, aka validation subset, is used to calculate its error rate to determine the convergence or stopping point. The training set is split into two subsets using stratified sampling: 90% of the training set is extracted in order to train the proposed model and the remaining 10% is kept to calculate the model's error. Each stratum is formed by ten different classes following a frequency distribution. In other words, the samples are picked randomly from each attack class as well as Normal records.Next, the training subset is processed with the Elastic Net feature selection algorithm. This algorithm selects 24 features out of 40 before the training subset is used to train the Balanced Bagging and XGBoost classifiers. Concurrently, the SFS algorithm selects 8 informative features out of 24 that were already selected among the original 40 by the Elastic Net. The output of SFS, which are 8 features, is used to train the Random Forest classifier. The classifier development schematic diagram has shown in Fig. 2.

Each classifier generates a matrix consisting of probability scores. Entries in this matrix are computed by dividing the number of votes for each class by the number of decision trees in each model. For instance, if the model had 30 decision trees in Random Forest and 20 of them vote for normal class on a new sample, the probability of Normal class is 0.67 (20/30). The first seven rows of the matrix produced by the Balanced Bagging is shown in Table III. It consists of 10 columns, representing 10 (9 attack plus Normal) classes, and N rows, where N designates the number of data samples in the validation subset. Since the validation subset has 35,068 records representing 10% of the training set, each model generates a probability matrix consisting of 35,068 rows. Both the probability matrix and the confusion matrix produced by XGBoost and Balanced Bagging classifiers are used as the inputs of Algorithm #1. This algorithm is employed to process the XGBoost and Balanced Bagging outputs to compute the errors caused by class overlap issue associated with the dataset.

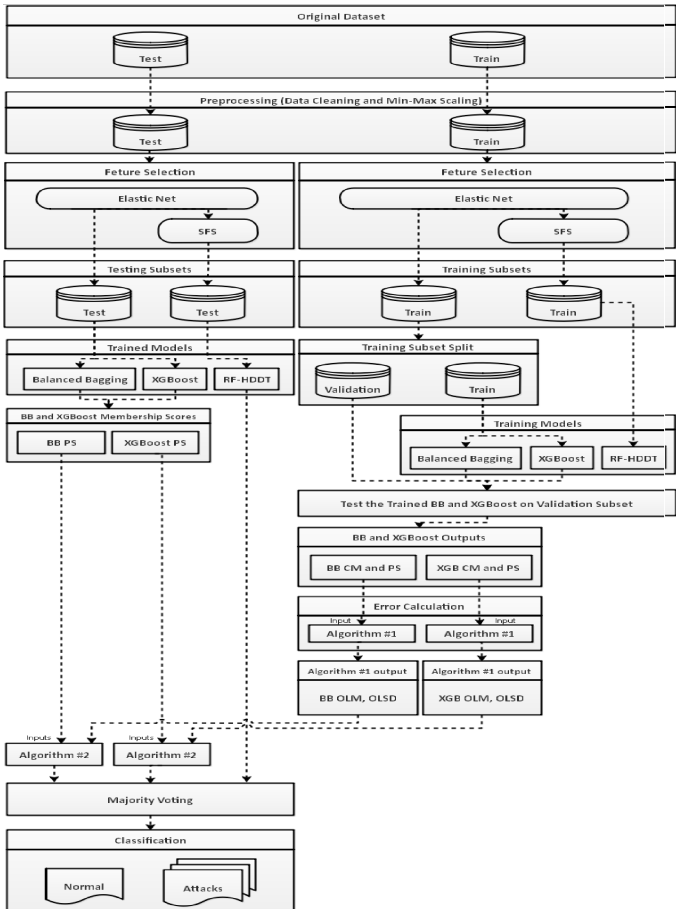


Fig. 2. Classifier Development Schematic Diagram

The output of Algorithm #1 is a nested list in Fig. 3 that consists of 10 items representing all the existing attack classes along with the Normal class which constitute the Level-1. Each item in Level-1 points to 9 sub-items in Level-2 as well. The corresponding sub-items in Level-2 for each item will not hold the item itself in Level-1. For each sub-item in the Level-2 of the nested list, there is a corresponding two-element list in Level-3. To make it more clear, it can consider this sole nested list as two two- dimensional arrays with the same dimensionality as the confusion matrix (10 by 10) storing mean and standard deviation. In other words, one matrix could hold the mean and another would hold the standard deviation values. Each row and column represents nine different attack classes along with the Normal class with the same order, similar to the rows and columns of the confusion matrix. These arrays store zeros along their main diagonal. The reason for that is that the aim of Algorithm #1 along with Algorithm #2 is to find the prediction errors or the errors existing in the membership scores. Since the main diagonal is holding true positives in confusion matrix, there is no error to calculate. That is why in the nested list, these entries are eliminated automatically.

Investigating the confusion matrix using Algorithm #1, if class A is misclassified x different times as class B, this algorithm iterates x times to calculate the difference between

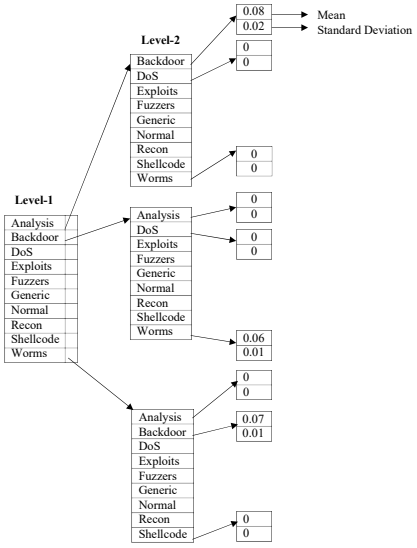


Fig. 3. Illustration of Partial Output by Algorithm #1

TABLE III. FIRST SEVEN ROWS OF A PROBABILITY SCORE MATRIX GENERATED BY BALANCED BAGGING CLASSIFIER

Row Index	Analysis	Backdoor	DoS	Exploits	Fuzzers	Generic	Normal	Recon	Shellcode	Worms
0	0.0083	0.0103	0.0988	0.1506	0.1718	0.0167	0.0750	0.0212	0.4463	0.0003
1	0.0034	0.0044	0.0632	0.1580	0.4279	0.0240	0.2884	0.0118	0.0144	0.0045
2	0.0057	0.0087	0.0641	0.1094	0.5057	0.0141	0.2659	0.0143	0.0111	0.0011
3	0.0062	0.0101	0.0627	0.1301	0.5491	0.0118	0.1968	0.0154	0.0159	0.0017
4	0.0018	0.0036	0.0339	0.0843	0.6023	0.0113	0.2427	0.0079	0.0109	0.0013
5	0.0044	0.0080	0.0827	0.1667	0.5055	0.0165	0.1213	0.0149	0.0774	0.0027
6	0.0034	0.0051	0.0605	0.1625	0.4867	0.0242	0.2271	0.0105	0.0140	0.0061

the probability score of class B and class A as well as class B and the eight remaining classes. If the former difference is smaller than the latter, this difference value (between class B and class A) is stored in a temporary variable (array D) for further calculation. Otherwise, the value is discarded. In the next step, the mean and standard deviation of the stored values are calculated and kept in arrays M and SD, respectively. These two values are placed in the third level of the nested list, which is an output of Algorithm #1, where the class A is the element of the first level of the list and class B is the element of the second level of the list.

To clarify how Algorithm #1 works, its application has been trace step by step next. The first row of Table IV(a) represents a confusion matrix for the Analysis attack and Table IV(b) represents the probability scores generated by the Balanced Bagging classifier in a two-dimensional array or matrix form after it is trained and its performance evaluated on the validation subset. Values of these matrices are held by CM and PS, two-dimensional array variables in the Algorithm #1, respectively. Initially, DL, OLM and OLSD are empty lists and eventually holding values for distances, output for mean values, and output for standard deviation values, respectively. AL is another list that initially contains the Normal and all the attack classes.

Algorithm 1 – Compute Means and Standard Deviations

Require: Mean-Standard-Deviation(CM, PS)
in:
two-dimensional array CM10×10 holding the confusion matrix and two-dimensional array PSn×10 holding the membership scores, n = the number of samples of validation subset
out:
two-dimensional array OLM10×10 initialized with zero
two-dimensional array OLSD10×10 initialized with zero
local:
empty array DL representing the minimum value of the membership scores difference and empty variable SD representing the computed Standard Deviation value and empty variable μ representing the computed Mean value and empty variable D representing the value obtained by subtracting the membership
scores constant:
array AL = {Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Normal, Recon, Shellcode, Worms}

```
1: for all x ∈ AL do ▷ AL is an ordered list
2:   for all y ∈ ( AL - x) do
3:     if CM_{x,y} > 0 then
4:       j ← 0
5:       for i ← 1 ... CM_{x,y} do
6:         D ← PS_{i,y} - PS_{i,y} ▷ where x is misclassified as y
7:         for all z ∈ (AL - (x, y)) do
8:           DT ← PS_{i,y} - PS_{i,z}
9:           if DT < D then
10:                count ← 1
11:            end if
12:          end for
13:          if count = 0 then
14:            DL_j ← D
15             j ← j + 1
16:          end if
17:        end for
18:        sum ← 0
19:        for h ← 1 ... length(DL) do
20:            sum ← sum + DL_h
21:        end for
22:        μ ← SUM/N
23:        SD ← sqrt((1/N) Σ_{i=1}^{N} (DL_i − μ)^2)
24:        OLM_{x,y} ← μ
25:        OLSD_{x,y} ← SD
26:      end if
27:    end for
28:  end for
29: return OLM, OLSD
```

TABLE IV. (A) CONFUSION MATRIX AND (B) FIRST TWO ROWS OF PROBABILITY SCORES MATRIX.

Class No.	Analysis (Class 0)	Backdoor (Class 1)	DoS (Class 2)	Exploits (Class 3)	Fuzzers (Class 4)	Generic (Class 5)	Normal (Class 6)	Recon (Class 7)	Shellcode (Class 8)	Worms (Class 9)
0	501	124	0	0	46	0	3	3	0	0
1	0	302	0	0	0	12	0	0	0	1
2	3	0	270	4	0	0	0	0	5	0
3	0	0	0	254	8	0	50	0	0	0
4	3	0	0	32	345	0	23	0	0	0
5	1	0	0	0	9	138	0	0	0	0
6	0	0	0	54	78	0	876	0	0	0
7	0	0	0	0	0	0	8	132	0	0
8	0	0	0	17	0	0	0	1	187	0
9	1	0	0	1	0	0	0	0	2	74

(A)

Index	Analysis	Backdoor	DoS	Exploits	Fuzzers	Generic	Normal	Recon	Shellcode	Worms
1	0.78	0.81	0.02	0.01	0.24	0.08	0.11	0.19	0.08	0.22
2	0.09	0.54	0.01	0.03	0.41	0.04	0.01	0.06	0.12	0.14

(B)

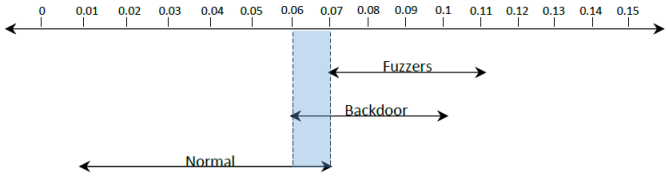


Fig. 4. An Example to Show how Algorithm 1 Calculates the Prediction Error Range

TABLE V. (A) THE CONFUSION MATRIX, (B) MEAN AND STANDARD DEVIATION ARRAYS THROUGH ALGORITHM #1 AND (C) MEAN AND STANDARD DEVIATION ARRAYS THROUGH ALGORITHM #2

	Analysis	Backdoor	DoS	Exploits	Fuzzers	Generic	Normal	Recon	Shellcode	Worms
Analysis	501	124	0	0	46	0	3	3	0	0

(a)

Analysis	Mean	0.08	0	0	0.09	0	0.04	1.3	0	0
	SD	0.02	0	0	0.02	0	0.03	0.1	0	0

(b)

Analysis	Mean	0.08	0	0	0	0	0	1.3	0	0
	SD	0.02	0	0	0	0	0	0.1	0	0

(c)

Fig. 4 shows that the Fuzzers, Backdoor, and Normal overlap. In this figure, Fuzzers represents a range of value between 0.09 -0.02= 0.07 and 0.09+0.02=0.11, Backdoor has a range of values between 0.08 -0.02= 0.06 and 0.08 +0.02= 0.1, and Normal is associated with a range of values between 0.04 - 0.03= 0.01 and 0.04+ 0.03= 0.07 where Analysis is incorrectly predicted as Fuzzers, Backdoor, and Normal, respectively. These values are taken from Table V(b) and after finding the overlaps, Table V(c) would be the final Mean and SD values generated for all the classes in preparation for the test phase. In Table V(c) the Mean and SD values for Recon remains unchanged since its error range is from 1.3-0.1= 1.2 to 1.3+0.1=1.4 and does not have any overlap with other ranges. Since Fuzzers, Backdoor, and Normal ranges overlap as well as the greater number of Analysis records are misclassified as Backdoor, Mean and SD values calculated for Backdoor remain unaltered and the Mean and SD values associated with Fuzzers and Normal are changed to zero. The process of finding the range overlaps and addressing them takes place for all of 10 classesThis procedure gives us a list of Means and SDs that is shown in Fig. 3. The list is eventually utilized in the test phase to minimize the prediction errors caused by data overlap.

D. Testing Methodology

In this phase, Mean and SD values computed by Algorithm #1 and revised after finding the overlaps are used to reduce the errors made by XGBoost and Balanced Bagging classifiers in identifying unseen samples. Due to data overlap issue associated with UNSW-NB15, classifiers may tend to predict class membership for certain new samples incorrectly. A new sample mimicking the behavior of data points belonging to another attack class is most likely to be misclassified. The probability scores matrix generated by the classifiers for a new sample in class 'A' contains error if this sample is incorrectly classified as class 'B'. In this case, the probability matrix score for class 'A' is lower than that of class 'B', while it must be just the opposite. The model consider the difference between the probability scores for classes 'A' and 'B' as error.

Algorithm 2 Membership Score Modification

Require: *Membership-Score-Modification(PS, OLM, OLSD)*
in: two-dimensional array PSn×10 holding the membership scores, n = the number of samples of test subset two-dimensional array OLM10×10 two-dimensional array OLSD10×10
out: two-dimensional array PSn×10 holding the (modified) membership scores, n = the number of samples of test subset
1: **for** i ← 1 … n **do**
2: max ← 0 ▷ holding zero in max to find the maximum membership score from line 3 to 8
3: **for** j ← 1 … 10 **do**
4: **if** $PS_{i,j} >$ max **then**
5: max ← $PS_{i,j}$
6: index_max ← j
7: **end if**
8: **end for**
9: min ← 10^{10} ▷ holding a very big constant value in min to find the minimum values obtaining from line 10 to 15
10: **for** j ← 1 … 10 **do**
11: **if** ((max-$PS_{i,j}$) < min) and (index_max ≠ j) **then**
12: min ← max - $PS_{i,j}$
13: index_min ← j
14: **end if**
15: **end for**
16. **if**(min≥($OLM_{index\_min,index\_max}$ - $OLSD_{index\_min,index\_max}$)) and (min≤($OLM_{index\_min,index\_max}$+$OLSD_{index\_min,index\_max}$)) **then**
17: $PS_{i,index\_min}$ ← ($PS_{i,index\_min}$ + $OLM_{index\_min,index\_max}$)
18: **end if**
19: **end for**
20: **return** PS

To reduce this error, Algorithm #2 has been applied on the probability score matrices generated through XGBoost and Balanced Bagging classifiers.

Since only the test set is utilized to evaluate the performance of our model, it is definitely unknown to the model. So, the model does not know the real target variable and it cannot calculate the errors using ground truth. This algorithm is designed to go through the membership scores generated by XGBoost and Balanced Bagging classifiers for the test subset in order to calculate the errors regardless of real target variables. The following steps discuss the functionality of Algorithm #2:

Using the revised mean and standard deviation values obtained by implementing Algorithm #1, Algorithm #2 is able to realize and correct the errors in the probability matrices where the errors arise from data overlap. Although, errors are not zeroed out using this algorithm, they may be reduced appreciably. The modified membership scores were used along with the membership scores obtained by the augmented Random Forest in the testing phase to make the final prediction using the majority vote. Using this method, the most voted prediction wins and taken as a final prediction. In other words, if more than two classifiers cast a vote for a particular class, the class will gain the final vote.

IV. Experiment and Result Analysis

An Intel Core i7 processor with 16 GB of RAM and Python with TensorFlow was used to create the proposed system. Before getting into the results, it is necessary to provide some background information on the UNSW-NB15 dataset. This dataset has been selected because of its advantages over older standard datasets. The lack of current cyberattacks types in the KDD98, KDDCUP99, and NSLKDD datasets and insufficient normal traffic and an unbalanced distribution of classes in the training and testing sets cause the datasets to suffer. The UNSW-NB15 benchmark dataset, explicitly designed for IDS design, has been presented to address these issues.

A. Experimental Results Evaluation

The confusion matrix is used to calculate performance measures for classifiers. The confusion matrix is used to calculate True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP, TN, FP, and FN denote records that are correctly identified as positives, negatives, or incorrectly identified as positives. The most standard measures include sensitivity, specificity, false positive, false negative, precision, and accuracy [22]. Abnormalities in a dataset must be taken into account while evaluating it. Due to the imbalanced datasets, accuracy is not appropriate. Insufficient datasets can use F-measure. This study aims to evaluate accuracy using the same parameter as previous studies has shown in equation (3). Accuracy is identified as the ratio of the correct classifications to the total number of samples and defined by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Sensitivity or Detection Rate (DR), shown in equation (4), corresponds to the proportion of true positives to all positives. It measures the probability of a sample being actually positive from all positive data points. Specificity indicates the proportion of false positives to all negatives. It measures the probability of a sample being actually positive from all data points that are predicted to be positive. They are used when the only positive or negative matter.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

False Positive Rate (FPR) or False Alarm Rate (FAR) represents the ratio of incorrect positive predictions to the overall number of negatives. At the same time, the Precision measures the probability of samples classified as positives for actually being positive. These two metrics are defined as shown in equation (5) and (6).

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

False Negative Rate (FNR) or Missed Alarm Rate, as shown in equation (7), indicates the ratio of incorrect negative predictions to the total number of positives. The evaluation metrics used in this study are based on the parts of the confusion matrix.

$$FNR = \frac{FN}{FN + TP} \tag{7}$$

F-measure is the harmonic mean of Sensitivity and Precision and is given by equation (8).

$$F - Measure = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{8}$$

B. Binary Classification

Table VI shows that 790 attack records are misclassified among the entire 45,332 attack records. It means that less than 2% of the overall attacks are misclassified as non-attack or Normal which leads to 0.017 missed alarm rate. In other words, the sensitivity for the attack class is 98.26%. On the other hand, 1022 of 37,000 Normal records are incorrectly classified as attacks which indicates less than 3% false alarm rate. Several evaluation metric values are shown in Table VII in order to comprehensively assess the performance of the proposed classifier design.

TABLE VI. Confusion Matrix on Test Data Subset for Binary Classification

PredictedActual	Normal	Attack
Normal	35978	1022
Attack	790	44542

TABLE VII. Performance Evaluation of the Proposed Model for Binary Classification

Metrics Class Type	Sen (%)	Spe (%)	FPR	FNR	Precision (%)	F-measure (%)
Normal	97.24	98.26	0.017	0.028	97.85	97.75
Attack	98.26	97.24	0.028	0.017	97.76	98.01

The main objective of any intrusion detection system (IDS) is to identify the pattern of the network traffic that may imply a suspicious activity. Accordingly, the performance of proposed IDS on UNSW-NB15 data is competitive in comparison with the other studies reported in the literature as shown in Table VIII. Two columns are dedicated to present the performance of the proposed classifier design. The first column indicates the average of calculated metrics in Table VII for both the attack and Normal records. On the other hand, the second column shows the values of performance metrics associated with the attack class in Table VII, which suggests that the performance of the proposed classifier design.

TABLE VIII. COMPARISON OF THE PROPOSED CLASSIFIER DESIGN WITH FOUR OTHERS CITED (NR INDICATES NOT REPORTED)

Metrics	Proposed Design (Avg)	Proposed Design	[10] (Avg)	[11]	[12]	[13]	[14]	[15]	[16]
Sensitivity	97.75	98.26	79.12	85	NR	98.47	NR	NR	NR
FNR (%)	2.25	1.74	NR	15	NR	NR	NR	NR	NR
FPR (%)	2.25	2.76	NR	2	8.6	2.18	NR	5.56	NR
Precision (%)	97.81	97.76	NR	99	NR	NR	NR	NR	96
F-measure (%)	97.88	98.01	77.87	91	NR	NR	NR	NR	95
Accuracy (%)	97.8	97.8	NR	89	91.31	94.11	81.42	86.49	95

C. Multiclass Classification

In this research, the performance of different estimators were evaluated separately by implementing them on the refined dataset. The results are shown in Table I. This study combined the first three estimators produced satisfactory performances to form an ensemble method along with Elastic Net and Sequential Forward Selection while Min-Max scaler had already implemented on the refined dataset. The results of the model evaluation is shown in Table IX in terms of confusion matrix. Although the outcome of the ensemble method has effectively improved the performance of each classifier contributed in the method, the model still suffered from generating large number of false negatives. In order to cover this issue, the study proposed two algorithms utilizing the basic statistic methods such as mean and standard deviation. Comparing Table IX with the performance of the proposed model depicted in detail for the multi-class case in Table X, represents the significant improves in most classes, such as Normal class to shrink the number of false negatives. This improvement verifies the effectiveness of Algorithm #1 and Algorithm #2. Although the study observe the increasing number of false negatives in some elements in confusion matrix, such as Fuzzers incorrectly classified as Backdoor, when the proposed algorithms implemented on the final decision, they are mostly seen between two attack class rather than an attack class and Normal records. On the other words, false negatives in one attack class may be increase due to the attack class to attack class misclassification. In Table X, the share of attack records which are incorrectly categorized as Normal traces is 4.14% for the 28% overall missed alarm rate. The remaining 23.86% missed alarm rate is associated with misclassification among attack classes which is not equally as problematic for network transactions. Although, 4.14% missed alarm rate for attack records alone could be detrimental for an intrusion detection system, our design achieves better results in comparison with the classifiers in other studies as shown in Table XI. Normal traces are also misclassified as Shellcode, Fuzzers and Analysis which suggests approximately 3% false alarm rate and 97.24% sensitivity.

TABLE IX. CONFUSION MATRIX SHOWING THE PERFORMANCE OF THE ENSEMBLE METHOD

Class No.	Analysis (Class 0)	Backdoor (Class 1)	DoS (Class 2)	Exploits (Class 3)	Fuzzers (Class 4)	Generic (Class 5)	Normal (Class 6)	Recon (Class 7)	Shellcode (Class 8)	Worms (Class 9)
0	327	193	4	1	18	0	130	0	4	0
1	298	154	8	8	23	0	76	1	15	0
2	1477	737	775	435	221	0	228	45	164	7
3	1322	1583	51	6197	257	0	667	490	420	145
4	651	386	8	14	1837	0	2604	7	540	15
5	15	30	37	390	99	18145	63	7	72	13
6	842	1	3	28	1427	0	34545	0	154	0
7	97	210	0	35	14	0	31	2967	128	14
8	11	0	0	0	7	0	9	3	347	1
9	0	0	0	1	0	0	1	0	3	39

TABLE X. CONFUSION MATRIX SHOWING THE PERFORMANCE OF THE PROPOSED DESIGN

Class No.	Analysis (Class 0)	Backdoor (Class 1)	DoS (Class 2)	Exploits (Class 3)	Fuzzers (Class 4)	Generic (Class 5)	Normal (Class 6)	Recon (Class 7)	Shellcode (Class 8)	Worms (Class 9)
0	327	193	4	1	18	0	130	0	4	0
1	126	405	0	0	13	0	39	0	0	0
2	1604	625	1110	228	92	0	178	59	164	29
3	779	830	39	8511	57	0	338	221	213	144
4	482	491	4	34	4380	0	0	10	646	15
5	15	30	37	390	99	18145	63	7	72	13
6	200	0	0	0	668	0	35978	0	154	0
7	103	207	0	35	14	0	31	2967	126	13
8	0	0	0	0	7	0	9	3	358	1
9	0	0	0	1	0	0	2	0	5	36

This is because these attack types mimic the behavior of Normal records [20, 23, 24]. This is the main reason that some attacks are also incorrectly predicted as Normal activity. In the associated confusion matrix, it can see that 19.20% of Analysis, 6.69% of Backdoor, 4.35% of DoS, 3.04% of Exploits, 0.33% of Generic, 0.88% of Reconnaissance, 2.38% of Shellcode, and 4.55% of Worms attack records are confused with Normal records.

Performance comparison of the proposed model with those studies reported in the literature is presented in Table XII and Table XIII. Many of the relevant performance metrics including the missed alarm rate, which is one of the most critical ones, are not reported in these studies by others. Consequently, performance comparison is done only for accuracy and sensitivity as these are the only metrics commonly reported in the cited studies.

Fig. 5 depicts the performance of the model in comparison with two models, Integrated and Dendron [18,19] that are proposed recently in terms of the F-measure. The proposed model in this study outperforms the other two given the F-measure values. The main reason is that the imbalance and overlapping problems in our model are addressed. This

TABLE XI. PERFORMANCE OF THE PROPOSED DESIGN FOR MULTI-CLASS CASE

Metric / Class	Accuracy	Sensitivity	Specifity	FPR	FNR	Precision	F-measure
Class0	95.56	48.30	95.95	0.041	0.51	0.15	64.25
Class1	96.89	69.47	97.09	0.029	0.31	0.84	80.99
Class2	96.27	27.15	99.89	0.001	0.73	0.42	42.70
Class3	95.98	76.46	99.03	0.009	0.24	0.84	86.29
Class4	96.78	72.25	98.73	0.012	0.28	0.77	83.44
Class5	99.12	96.15	100.0	0.000	0.28	1.00	98.04
Class6	97.81	97.24	98.26	0.017	0.02	0.97	97.75
Class7	95.39	84.87	95.86	0.041	0.15	0.61	90.03
Class8	99.31	94.71	98.33	0.017	0.05	0.34	96.49
Class9	99.73	81.82	99.74	0.003	0.18	0.24	89.90

CLASS0: ANALYSIS CLASS1: BACKDOOR CLASS2: DOS CLASS3: EXPLOITS CLASS4: FUZZERS CLASS5: GENERIC CLASS6: NORMAL CLASS7: RECONNAISSANCE CLASS8: SHELLCODE CLASS9: WORMS

TABLE XII. THE ACCURACY OF PROPOSED MODEL VS. THE ACCURACY OF DIFFERENT MODELS (NR: NOT REPORTED)

Attack Type	Accuracy	Accuracy [17]	Accuracy [18]	Accuracy [19]	Accuracy [20]	Difference
Analysis	95.56	**99.44**	99.26	99.3	NR	-3.88
Backdoor	96.89	99.06	**99.11**	97.93	NR	-2.22
Dos	**96.27**	96.14	94.9	95.71	94.52	0.11
Exploits	**95.98**	93.91	90.12	93.58	89.72	2.07
Fuzzers	**96.78**	96.52	91.47	95.04	NR	1.74
Generic	**99.12**	98.34	98.23	98.7	87.7	0.26
Normal	97.81	98.16	93.54	94.59	**98.64**	-0.3
Reconnaissance	95.39	98.74	95.33	96.18	**99.1**	-3.71
Shellcode	99.31	99.22	**99.4**	98.33	NR	-0.09
Worms	99.73	97.28	**99.92**	99.78	NR	-0.14

TABLE XIII. THE SENSITIVITY OF PROPOSED MODEL VS. THE SENSITIVITY OF DIFFERENT MODELS (NR: NOT REPORTED)

Attack Type	Sensitivity	Sensitivity [19]	Sensitivity [20]	Difference
Analysis	48.30	20.45	NR	+27.85
Backdoor	69.47	67.32	NR	+02.15
Dos	27.15	14.29	5.0	+12.86
Exploits	76.46	76.22	54.64	+00.24
Fuzzers	72.25	64.42	NR	+07.83
Generic	96.15	81.37	96.72	-00.57
Normal	97.24	97.39	98.00	-00.76
Reconnaissance	84.87	46.04	71.70	+13.17
Shellcode	94.71	36.39	NR	+58.32
Worms	81.82	18.37	NR	+63.45

work have a combination of ensemble methods to handle the imbalance and if-then-else rules to mitigate the adverse effects of overlapping issue and using Hellinger distance criterion to choose the best split considering the imbalance problem.

V. CONCLUSION

This study presents design and performance evaluation of an intrusion detection (and identification) system using machine learning for the UNSW-NB15 dataset. The study evaluated the performance of classifier design which employs three ensemble classifiers and two proposed algorithms where the later is developed for minimizing the errors due to one of the two issues inherent to the UNSW-NB15 dataset, namely the class overlap and class imbalance. To deal with the imbalanced data, this work utilized the Balanced Bagging and the XGBoost ensemble classifiers which offer a set of hyper-parameters that,

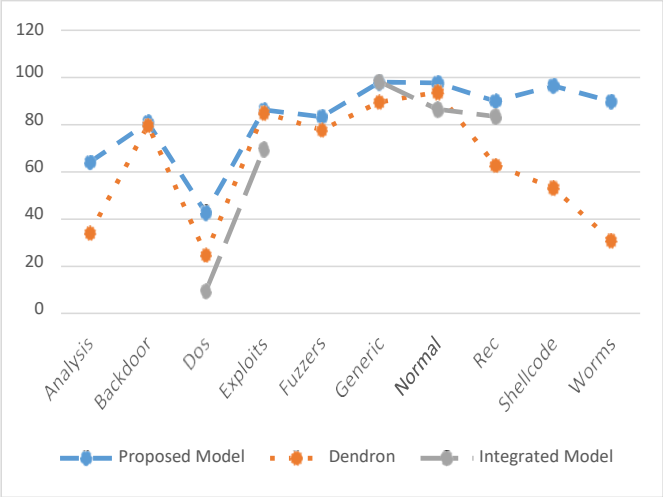


Fig. 5. F-measure Comparison

through judicious adjustments of the same, help contribute to improved performance in the presence of the imbalanced data. To address the class overlap issue, the study proposed two algorithms and utilized them to process and modify the classification outputs from the Balanced Bagging and the XGBoost ensembles. Outputs of three ensemble classifiers, namely Random Forest, Balanced Bagging and XGBoost, were provided as inputs to a combiner that implemented majority voting to determine the final class membership of an input data record under test. The performances of the classifiers are assessed by employing six different normalization methods on the modified UNSW-NB15. The results showed that min-max scaler enhanced the performance of the classifiers in terms of accuracy. Min-max scaler as a normalization method helped increase the distances between the data points of two different attack classes reducing the degree of class overlap. Application of the combination of preprocessing, feature selectors, tree-based ensemble classifiers, and the proposed algorithms for this design resulted in superior performance when compared to seven other classifiers, reported in the recent literature, implemented on the UNSW-NB15 dataset for both multi-class and binary classification cases. Performance of the proposed model was compared with both the binary classifiers and multi-class classifiers cited in the literature. In the binary class classification case, this model could classify more than 98% of the attack classes correctly. The model also performed highly for the classification of Normal records with more than 97%.

In comparison with other studies reported in the current literature on the UNSW-NB15 dataset, this model achieved impressive results. It addresses two major issues that a dataset may suffer from, overlap and imbalance. The study employed Balanced Bagging and XGBoost offering a range of hyperparameters in order to address the dataset imbalance. Also, the study utilized the Hellinger distance for the Random Forest for the same reason. The study further proposed two new post-processing algorithms for the outputs of training models to minimize the errors caused by the large number of impure nodes generted during the training phase due to the data overlap issue. In future this work plan to use Hellinger distance

as split criterion for both Balanced Bagging and XGBoost to enhance the performance of this model. The future works are also aimed at utilizing the proposed algorithms, known as Algorithm #1 and Algorithm #2, along with Random Forest.

REFERENCES

[1] A. K. Pandey, A. K. Tripathi, G. Kapil, V. Singh, M. W. Khan, A. Agrawal, et al., "Trends in Malware Attacks: Identification and Mitigation Strategies," in Critical Concepts, Standards, and Techniques in Cyber Forensics, ed: IGI Global, 2020, pp. 47-60.

[2] J. L. Cebula and L. R. Young, "A taxonomy of operational cyber security risks," Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst2010.

[3] E. C. Thompson, Cybersecurity Incident Response: How to Contain, Eradicate, and Recover from Incidents: Apress, 2018.

[4] S. Morgan, "Official annual cybercrime report," Sausalito: Cybersecurity Ventures,2019.

[5] J. Armin, B. Thompson, D. Ariu, G. Giacinto, F. Roli, and P. Kijewski, "2020 cybercrime economic costs: No measure no solution," in 2015 10th International Conference on Availability, Reliability and Security, 2015, pp. 701-710.

[6] Uikey, R.; Gyanchandani, M.: Survey on classification techniques applied to intrusion detection system and its comparative analysis. Int. Conf. Commun. Electron. Syst. 2019, 1451–1456 (2019)

[7] Soe, Y.N.; Feng, Y.; Santosa, P.I.; Hartanto, R.; Sakurai, K.: Machine learning-based iot-botnet attack detection with sequential architecture. Sensors 20(16), 4372 (2020)

[8] Ahmad, Z.; ShahidKhan, A.;Wai Shiang, C.;Abdullah, J.;Ahmad, F.: Network intrusion detection system: a systematic study of machine learning and deep learning approaches. Trans. Emerg. Telecommun. Technol. 32(1), e4150 (2021)

[9] Chebrolu, S.; Abraham, A.; Thomas, J.P. Feature deduction and ensemble design of intrusion detection systems. Comput. Secur. 2005, 24, 295–307. [CrossRef]

[10] V. Kumar, A. K. Das, and D. Sinha, "UIDS: a unified intrusion detection system for IoT environment," Evolutionary Intelligence, pp. 1-13, 2019.

[11] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset," Cluster Computing, pp. 1-22, 2019.

[12] B. A. Tama and K.-H. Rhee, "An in-depth experimental study of anomaly detection using gradient boosted machine," Neural Computing and Applications, vol. 31, pp. 955-965, 2019.

[13] M. G. Raman, N. Somu, S. Jagarapu, T. Manghnani, T. Selvam, K. Krithivasan, et al., "An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm," Artificial Intelligence Review, pp. 1-32, 2019.

[14] Khammassi, Chaouki and Saoussen Krichen. "A GA-LR wrapper approach for feature selection in network intrusion detection." Comput. Secur. 70 (2017): 255-277.

[15] Tian, Q., Han, D., Li, KC. et al. An intrusion detection approach based on improved deep belief network. Appl Intell 50, 3162–3178 (2020). https://doi.org/10.1007/s10489-020-01694-4

[16] Abirami, M. & Yash, Umaretiya & Singh, Sonal. (2020). Building an Ensemble Learning Based Algorithm for Improving Intrusion Detection System. 10.1007/978-981-15-0199-9_55.

[17] Soulaiman Moualla, Khaldoun Khorzom, Assef Jafar, "Improving the Performance of Machine Learning-Based Network Intrusion Detection Systems on the UNSW-NB15 Dataset", Computational Intelligence and Neuroscience, vol. 2021, Article ID 5557577, 13 pages, 2021. https://doi.org/10.1155/2021/5557577

[18] J. Sharma, C. Giri, O.-C. Granmo, and M. Goodwin, "Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation," EURASIP Journal on Information Security, vol. 2019, p. 15, 2019.

[19] D. Papamartzivanos, F. G. Mármol, and G. Kambourakis, "Dendron: Genetic trees driven rule induction for network intrusion detection systems," Future Generation Computer Systems, vol. 79, pp. 558-574, 2018.

[20] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset," Cluster Computing, pp. 1-22, 2019.

[21] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," Data Mining and Knowledge Discovery, vol. 24, pp. 136-158, 2012.

[22] Abhijit Das, S G Balakrishnan and Pramod, "Network Intrusion Detection System based on Generative Adversarial Network for Attack Detection" International Journal of Advanced Computer Science and Applications(IJACSA), 12(11), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0121186

[23] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Information Security Journal: A Global Perspective, vol. 25, pp. 18-31, 2016.

[24] Y. Yang, K. Zheng, C. Wu, X. Niu, and Y. Yang, "Building an effective intrusion detection system using the modified density peak clustering algorithm and deep belief networks," Applied Sciences, vol. 9, p. 238, 2019.

An Efficient Feature Selection Approach for Intrusion Detection System using Decision Tree

Abhijit Das[1]
Research Scholar, Dept. of CSE
PES Institute of Technology & Management
Affiliated to VTU, Shivamogga, India

Pramod[2]
Associate Professor, Dept. of ISE
PES Institute of Technology & Management
Affiliated to VTU, Shivamogga, India

Sunitha B S[3]
Associate Professor, Dept. of CSE
PESITM, Affiliated to VTU
Shivamogga, India

***Abstract*—The intrusion detection system has been widely studied and deployed by researchers for providing better security to computer networks. The increasing volume of attacks, combined with the rapid improvement of machine learning (ML) has made the collaboration of intrusion detection techniques with machine learning and deep learnings are a popular subject and a feasible approach for cyber threat protection. Machine learning usually involves the training process using huge sample data. Since the huge input data may cause a negative effect on the training and detection performance of the machine learning model, feature selection becomes a crucial technique to rule out the irrelevant and redundant features from the dataset. This study applied a feature selection approach for intrusion detection that incorporated state-of-the-art feature selection algorithms with attack characteristic feature to produce an optimized set of features for the machine learning algorithms, which was then used to train the machine learning model. CSECIC- IDS2018 dataset, the most recent benchmark dataset with a wide attack diversity and features have been used to create the efficient feature subset. The result of the experiment was produced using machine learning models with a decision tree classifier and analyzed with respect to the accuracy, precision, recall, and f1 score.**

***Keywords*—Intrusion detection; feature selections; decision tree; machine learning; cyber security**

I. Introduction

Nowadays, computer networks have been applied to every aspect of people's lives and daily production. People use desktop computers, laptop computers, or other types of Internet enabled devices to access public information that has been published online. Different organizations, such as companies and schools, build up networks to exchange information within the organizations. Network security has become one of the major concerns of most people and organizations when using computers and other Internet-enabled devices to access online resources and store valuable data [1].

IDSs are generally used as an effective tool to defend network protection by identifying network attacks from malicious users on the Internet. These techniques examine traffic by analyzing the packet information at various layers of the communication model [2], a method known as packet analysis. The use of machine learning approaches to enhance IDS and solve network threats has risen in recent years. More and more researchers are beginning to employ machine learning approaches to detect and classify abnormal activities by allowing the method to learn various threats from example data [3]. Machine learning and deep learning techniques are becoming more sophisticated and used in various technological fields. Additionally, machine learning, which is a subset of artificial intelligence, has significant potential in cybersecurity.

The general process of machine learning could be straightforward and comprehensible, but it is never limited to what will be presented in this research. When researchers apply the machine learning approaches, they need to identify the features of sample data that are summarized patterns ofthe sample data in the packet of network traffic. For example, IP address, protocol, port number, etc. The researchers use features to train models, reach a higher identification rate and accuracy, and then apply the trained machine learning models to detect further and classify network attacks in network traffic. However, when using machine learning to study different problems in the same field, the features used for training and testing may not necessarily be the same. For instance, detecting DDoS attacks may require investigating different packet information from detecting spam. Hence, solid domain knowledge and choosing the right features play an important role in machine learning.

Applying machine learning in network intrusion detection is well studied. Many studies on machine learning in NIDS are published each year. Weirdly, researchers seem to be eager to experiment with various machine learning techniques for detecting network intrusion. Yet, at the same time, many investigators are stingy at explaining why they choose certain features or use up all features in the dataset, except for some researchers who are focusing on the feature engineering aspect of machine learning.

After taking a deeper look at machine learning, it is easy to notice that using a large number of non-representative features could create the challenging problem for creating an effective and accurate ML model. Feeding a large amount of data will create millions of possibilities for a machine learning model, making it hard to distinguish the attack patterns during the monitoring process [4]. On the other hand, the redundant or irrelevant feature is one of the most critical factors that force excessive training and classification time [5]. However, the importance of feature selection is often neglected or underrepresented by some researchers [6]. Attribute selection plays an essential role in creating the machine learning model for classifying or predicting purposes. Each kind of network attack has specific attack patterns that could be discovered in the data sorted in different features [7]. In many types of research about machine learning and deep learning in network security, the significance of feature selection solutions has

been emphasized repeatedly. Applying the advanced feature selection approaches could significantly improve network intrusion detection systems; Cai et al. [8] found that effective attribute selection outcomes might enhance learning precision, decrease training time, and clarify results. A machine learning strategy that relies on interconnected essential features can cut down on the number of iterations of an experiment [9]. Over-fitting and model generalisation can be reduced by finding the best feature subset, which can assist reduce the number of features utilised for training machine learning models [10]. It is also much faster to process and train models with fewer data when fewer characteristics are fed into them [11]. The classification accuracy of an ML model can be improved by removing irrelevant features using feature selection techniques [12].

This work aims to design a new feature selection approach to enhance the machine learning models in network intrusion detection by creating the optimal subset of features. By finishing this study, multiple deliverables would be presented as follows. The importance of different features in network intrusion detection would be identified based on the characteristics of each kind of network attack. The optimal combinations of features for each network attack in the study would be identified by comparing the detection rate of different combinations. As a part of this work, data analysis will be carried out to demonstrate how the ML model's performance was improved utilising the best possible set of features.

Research Question:

- Is it possible for a machine learning model to produce better predictions for network intrusion detection using a hybrid method that combines feature selection techniques with attack characteristic features?

II. Related Work

It is possible to choose the most closely related features from a dataset without using machine learning techniques using a filter approach of feature selection [13]. Test scores from various statistical methodologies are all that is needed to determine the relationship between attributes. Linear or non-linear associations can exist between any of these numerous features. Many popular statistical procedures, including correlation coefficients and Chi-square tests, as well as the ANOVA test, are used. According to the statistical techniques, the attributes are ranked according to their correlation or joint distribution. In general, the more closely two features are correlated, the more closely they are linked; conversely, the less closely two features are correlated, the less closely these two features are related. Since the filter method is not dependent on any other complicated mining or validating methods, it is a simple implementation that effectively eliminates extraneous features.

The filter method of feature selection is commonly used in the data preprocessing stage. The machine learning model has not been applied yet with the sample data of the selected features that are decided in the feature selection stage. This characteristic creates another important advantage of the filter method, which makes building a machine learning model much faster. The features were only selected once using the filter method to create the subset of highly related features. Since the data dimension is reduced dramatically before the data is fed into the machine learning algorithm, it is less prone to over-fitting [14].

The wrapper feature selection method shows a significant difference from the filter method. The wrapper method evaluates the goodness of the features by considering the prediction results through the machine learning models [15]. The wrapper method's commonly used machine learning algorithms including SVM, DT, BN, k-means, RF, etc. To assess the accuracy and precision of every group of extracted features, such as the rate at which estimates are correct or incorrect, the machine learning model's training procedure must be repeated many times.

When using the wrapper technique, all possible combinations of features are tested to determine which set has the best accuracy and error rates. Over-fitting might slow down and complicate the wrapping process. In addition, the wrapper technique is less transitive because of the changes in ML concepts [16]. To ensure that the chosen features are compatible with the newly learned learning algorithm, it should be done again if the ML algorithm is modified after the learning process has concluded.

There are three major techniques in the wrapper feature selection method: forward feature selection, backward feature elimination, and Bi-directional elimination. Forward feature selection initiates selecting process when there is no feature in the feature subset, and a new feature that could best improve the prediction results of the machine learning model is added into the feature subset in each selecting iteration until the result cannot be improved anymore [17]. The features selected by this method represent the best subset of features that could achieve the highest accuracy rate during the classification. Backward feature selection is completely opposite to the forward selection, which starts the selecting process with all of the features in the dataset and removes one feature in every iteration that makes the largest decrease in the model's accuracy. The reducing process would repeat until the accuracy could not be further improved or all of the features have been exhausted [18]. Bi-directional elimination can be seemed as combining the forward selection with the backward elimination. This method first sets thresholds of significance level for the forward selection as well as the backward elimination. Then the forward selection is applied by adding one feature and examining the significance level of the feature in each selection round. After the forward selection has been finished, the backward elimination will be performed by removing the feature with a higher significance level than the elimination threshold in each reducing round. These two methods will be repeated until the optimal feature subset is found [19].

The Embedded Method of attribute Preference overcomes the disadvantages of filter and wrapper methods [20]. Because it is integrated into the learning process rather than being separate, the embedded approach makes it possible to pick features during the training process of ML algorithms and decreases data volumes internally. On the other hand, the embedded method is less likely to require many computing resources. The embedded technique has a better overall detection rate than the filter method because it interacts with the machine learning algorithms instead of relying purely on the rank of features.

For example, LASSO is a regression technique, while Decision Tree(DT) and Random Forest (RF) are examples of tree-based algorithms [21]. The embedded method optimizes the objective functions with the regularization penalty terms and measures the feature importance [22]. L1 regularization used by LASSO regression penalizes the weight of less important features to zero. The features that have the least coefficient will be eliminated automatically for achieving better detection results. The tree-based algorithms examine the feature importance. The features are permuted based on the importance score, and the most important feature will keep close to the tree's root.

Dataset is a key component of machine learning and is used to train the machine learning algorithms. The dataset contains all the information that machine learning may use to recognize and classify the patterns of the objective, for example, the network activities for intrusion detection. Dataset is composed of various features, and each represents a piece of data that indicates some information about the objective. Some of these features are directly extracted from the raw data, called basic features; for example, the IPs, port numbers and TCP flags are originally contained in the network packet. There are also a lot of features in the dataset that are the statistical information created by analyzing the raw data [23]. This kind of feature is called the derived feature. The researchers manually create these statistical features to describe the objective's characteristics better. Onut and Ghorbani divided the derived features into two major groups: single connection dependent and multiple connection dependent [24].

The single connection dependent derived features are created using the flow information from a single connection. The functionality of the single connection dependent features is to verify whether the current connection has malicious intent or not. The single connection dependent derived features could be used to detect the bursty and stealthy attacks based on the packet data within a certain time interval and the whole lifetime of a single connection. The examples of the single connection dependent derived feature include number of packets, packet length, number of TCP flags per packet, etc. The multiple connection dependent derived features are used to represent the relationships between multiple connections. These features are mostly used to detect any kind of network attack launched through multiple network connections, such as worm attacks, DDoS attacks, etc. In machine learning research in intrusion detection, some derived features are created to detect various network attacks better.

Najafabadi et al. introduced three derived features that present the packet information extracted from the network flow following the IPFIX standard to provide better detection of the brute-force attacks: the number of packets, packet size, initial flags, and session flags [25].

- The number of packets describes how many packets are captured in the flow. Since the attackers need to frequently guess the password of the users' accounts until they obtain the correct one, the packet number of brute force attacks would be much larger than that of normal login activities.
- Packet size describes the total size of the packets in the flow. On the reasoning of the small size of network flow for the failed logins, the normal login activities would have apparently larger byte size.
- Flow flags describe the flags of all packets seen in the flow. This feature can recognize the attack traffic containing the complete set of TCP flags, including FIN, SYN, PSH, and ACK flags, which is not normal for the common TCP connections with only SYN and ACK flags.

Constructed features mentioned above were applied in 5-Nearest Neighbor, C4.5 Decision Trees, and Naïve Bayes algorithms to predict SSH brute force attacks by Najafabadi et al. The results showed that the constructed feature significantly improved the performance of weak algorithms like Naïve Bayes and achieved 99% accuracy using 5-Nearest Neighbor and C4.5 Decision Trees.

The botnet attack can be detected using the derived features extracted from the network traffic mentioned in the last section. Since the bots need to contact the command and control server to obtain instructions for further malicious activities, TCP connections are required between bots and the command and control server. And this kind of TCP connection shows a periodic pattern according to the observations by Wang et al. [26]. Under this situation, the interval time between request and response flow could be an important feature. On the other hand, Wang et al. found that the TCP connections between bots and C&C server usually follow the periodical DNS query from bots to C&C server. Therefore, the information extracted from the DNS query can be used as an effective feature to detect botnet attacks. Three derived features based on the DNS query were introduced: interval time of DNS queries, the total number of DNS responses and the failed DNS responses [26]. As mentioned above, Jin et al. constructed similar features to predict the botnet attack using Adaboost, C4.5 Decision Trees, and Naïve Bayes algorithms. The result showed that using contracted features, Adaboost and C4.5 Decision Trees classifiers achieved over 90% for precision, recall, f1, and ROC area, and Naïve Bayes reached over 70% for all scores as well [27], which proved that the constructed features were critical for detecting botnet attack.

In order to pick up the web attacks more efficiently, eight derived features were introduced by Qin et al. using the information extracted from the webserver logs: diffReqPercent, stutas200Percent, avgBytePerRequest, urlLevelRate, maxFrequency, FrequencyTimes, requestTimeDistrubution, and avgInterval [28]. Each of these features can be computed using the statistical information based on the webserver records [29], including the total number of requests, different requests, and successful requests, and the total length of requests from the same user within a certain time interval. After applying the constructed features in Naïve Bayes, Radial Basis Function Network and C4.5 decision tree, Qin et al. achieved accuracy and detection rate for more than 98% and false positive rate for lower than 2%.

The port scanning attacks take advantage of the TCP, UDP, and ICMP responses to detect the accessible open ports of any hosts existing in the network. These scanning attacks usually get involved in the flows either to various ports in the single host or the same port in the multiple hosts. However, the network event associated with these protocols can also

be used to identify the port scanning attacks. Ring et al. created two derived features that target different kinds of port scanning attacks: ICMP-Error count and RST count [30]. Decision trees and Support Sector Machine were used to test the performance of constructed features. The results indicated that both classifiers reached 100% detection rate and 10% false alarm using the constructed features.

III. Methodology

Hypotheses for this study were as follows:

- H_0: The proposed feature selection approach of combining feature selection algorithms and attack characteristic features does not improve the detection performance when compared with the detection approach using all features in the given dataset.
- H_α: The proposed feature selection approach of combining feature selection algorithms and attack characteristic features improves the detection performance when compared with the detection approach using all features in the given dataset.

This study focused on Feature formation and feature selection in machine learning (ML) for identifying network threats. The study required a grasp of the ML process and development approach and mainly focused on investigating the suggested solution's effectiveness. The following is a step-by-step breakdown of the research process.

1) Understand the functionality and process of machine learning.
2) Study and analyze the current feature selection approaches for network intrusion detection.
3) Set up machine learning model.
4) Design a new feature selection approach and apply the proposed approach to obtain optimal feature subset.
5) Generate findings of the tests, analyze and document the performance of the proposed solution.

The machine learning model contained three major stages: preprocessing data, training model, and classifying target data. The detailed activities in each stage are depicted in Fig. 1.

1) Preprocessing data aimed to organize the raw training data in an acceptable data structure and convert the dataset into the proper format permitted by the machine learning models. Data preprocessing involves handling null values, categorical variables, standardization, one-hot encoding, and multicollinearity [31].
2) Training machine learning model allowed it to fit with the training data and tune model parameters for classification needs. This stage repeatedly adjusted the hyperparameters of the machine learning algorithm to find the function that best described the data.
3) Classifying target data was to apply the trained machine learning model to perform detection functionality on the test data that had never been fed into the model before.

The proposed approach of feature selection was divided into six steps shown in Fig. 2. **Firstly**, the CSE-CIC-IDS2018 dataset with raw features as input. **Secondly**, attack characteristic features that might indicate the attack patterns were added into the dataset. **Thirdly**, all data was preprocessed to make sure all features were processible by the machine learning model. **Fourthly**, all features were scaled to make the data normalized in magnitude. **Fifthly**, all features were calculated through multiple feature selection algorithms for filtering out the majority of irrelevant features. The **final step** was to output the attribute set to train the ML model and produce prediction results.

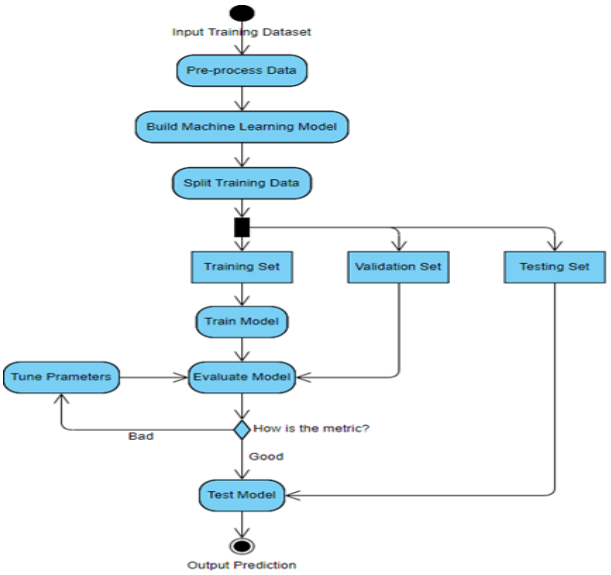


Fig. 1. Machine Learning Workflow

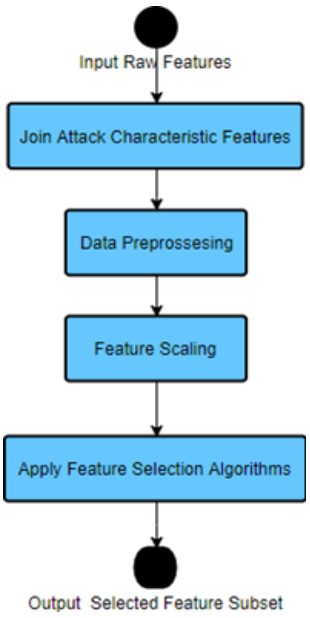


Fig. 2. Proposed Solution Process

The CSE-CIC-IDS2018 dataset contained benign background traffic and malicious traffic based on seven kinds of network attacks, including brute-force attack, Heartbleed

TABLE I. CONFUSION MATRIX

	Positive Prediction	Negative Prediction
Positive Condition	True Positive (TP)	False Negative (FN)
Negative Condition	False Positive (FP)	True Negative (TN)

attack, botnet attack, DoS attack, DDoS attack, web attacks, and infiltration attack. The attacks studied in this work were brute-force, botnet, web, and infiltration attacks. The dataset included seven features extracted from the raw data flow, for example, protocol, timestamp, IP address, etc.

This work evaluated the prediction results of different experiments using the test scores produced from the confusion matrix. The confusion matrix is a two-dimensional matrix that represents the correlation of true conditions and predictive results shown in Table I.

TP describes the number of abnormal samples being accurately classified. TN defines the number of normal samples being accurately classified. FP specifies the number of normal samples being falsely classified as abnormal samples. FN specifies the number of abnormal samples being falsely classified as normal samples. Various test scores were calculated using the confusion matrix in this work: Accuracy, Precision, Recall, and F1 Score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

IV. EXPERIMENTS

The experimental environment was implemented on JupyterLab hosting on a Jupyter Docker container that contained various Jupyter applications and interactive computing tools. All experiments were conducted on the server having 2.93 GHz 6 cores Intel Xeon X5670 CPU with 96 GB RAM for the physical device.

In this work, the selected machine learning classifier was a decision tree that continuously splits the data according to the defined parameters. The decision tree contains three major components: Nodes, Branch, and Leaves. The node represents a test for the data of a certain feature. The branch contains the result of a node and connects to the next node or leaf. The leaf is the final node of a tree that provides the prediction corresponding to the label of the sample data. The decision tree algorithm was specified with some parameters that helped to optimize the performance.

The decision tree employed the Gini impurity to evaluate the quality of data splitting. Gini impurity calculated the probability of a sample being randomly misclassified regarding the distribution of different labels. The algorithm also controlled the data selection to use different random values for each run of the classification by setting random_state to none. The decision tree classifier was set to use the best splitter that split the data on the most relevant feature instead of randomly shuffling the feature. Other hyperparameters of the selected architecture are shown in Table II.

TABLE II. SAMPLE HYPERPARAMETERS OF DECISION TREE ALGORITHM

Hyperparameter	Value
Criterion	gini
Random_state	None
Splitter	Best
Class_weight	None
Max_depth	None
Max_feature	None
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
Presort	False

The work involved three major stages: Preprocessing data, Training model, and Classifying target data. However, in order to present the technical details of model construction, when writing the code using Python programming language, the model was expended into six phases: Data Loading & Presentation, Data preprocessing, Feature Scaling, Feature Selection, Building the model, and Prediction & Evaluation. Table III summarizes the functions of all stages.

TABLE III. MODELLING STAGES AND DESCRIPTIONS

Stage	**Description**
Data Loading & Presentation	This stage was to import the CSE-CIC-IDS2018 dataset and provide the sample view and statistical summary of the dataset.
Data Preprocessing	This stage was to laundry the CSE-CIC-IDS2018 dataset.
Feature Scaling	This stage was to split train & test set and normalize the data of the CSE-CIC-IDS2018 dataset.
Feature Selection	This stage was to use the feature selection algorithms to produce a subset of features.
Building the model	This stage was to build the decision tree classifier.
Prediction & Evaluation	This stage was to predict the network attacks and evaluate the results of experiments.

A. Data Loading

The Data Loading & Presentation stage was to import the proper Python libraries and the CSE-CIC-IDS2018 dataset and provide the dataset's sample view and statistical summary. The CSE-CIC-IDS2018 intrusion detection dataset was released by the Communications Security Establishment (CSE). The updated version is structurally similar to CICIDS2017 and has a class imbalance as well. As a result, the dataset for CSE-CIC-IDS2018 contains 16,233,002 instances drawn from 10 days' worth of network traffic, as opposed to the smaller network used for CSE-CIC-IDS2018. Attack traffic accounts for about 17% of all occurrences. CSE-CIC-IDS2018 represents seven different types of network traffic. Ten CSV files containing the data are available for download from the cloud. There are 79 independent features in nine files and 83 independent features in the remaining nine files. Due to the scope of the project, network attacks related to Dos and DDOS weren't considered and used in the project. Therefore, the first step of Data Loading & Presentation stage was to remove the .csv

files containing data of Dos and DDOS attack from the dataset. All other csv files were required to import as a data-frame, a two-dimensional data structure containing labeled axes. Data-frame of the dataset was essential in order to convert the .csv file into Numpy array format containing all numerical values of dataset for further training the machine learning model.

B. Data Preprocessing

Laundering the CSE-CIC-2018 dataset and creating train and test sets were the objectives of this step. It was necessary to pre-process the dataset before feeding it to the machine learning model so that all values fit into the proper data type and were readable. The CSE-CIC-IDS2018 Dataset was cleaned using the following steps:

- Remove insignificant features.
- Create a standard data type for the data.
- Remove rows containing "Infinity" and "NaN" value.
- Reduce the long decimal digits of the float numbers.
- Rename attack labels.

C. Feature Scaling

The CSE-CIC-IDS2018 dataset was normalized and split into a train set and a test set at this stage. In order to create a train and test set for the CSE-CIC-IDS2018 dataset, a random percentage of the dataset had to be divided into two sets. The test set was a set of data that was never used in the train set and was employed to produce the prediction and evaluate the final machine learning model. In the project, the train and test sets were 80% and 20% of the processed CSE-CIC-IDS2018 datasets. Since some float numbers had extremely long digits after decimal, in order to avoid the memory issue with these long digits during the computing process, the number of digits after decimal was reduced to 1 for all float numbers. Label column contained label values for all instances. Labels were divided into eight categories: seven attack labels, mentioned in Data Loading section, and a benign label. However, string value couldn't be processed by machine learning model. Under this situation, all seven label values were renamed with numeric values from 0 to 7. Label column was taken up from the data-frame and replaced the label values with numbers respectively using labeldf.replace() function, then new Label column was put back to the data-frame.

This stage was to split the train set & test set and normalize the data of the CSE-CIC-IDS2018 dataset. After finishing the data laundry for the CSE-CIC-IDS2018 dataset, the next step was to randomly create a train set and test set by separating the dataset with a certain percentage. The train set was used to help the machine learning algorithms fit the parameters that best described the sample data. The test set was a set of data that was never used in the train set and was employed to produce the prediction and evaluate the final machine learning model. In the project, the train set and test sets were 80% and 20% of the processed CSE-CIC-IDS2018 dataset, respectively. The data frame was separated into x and y. x was assigned as a data frame of all features, and y was a series of labels in the original data frame. Then x and y were split into train set and test set respectively using *train_test_split()* function. The parameter *test_size* defined the proportion of the data-frame and series assigned to the test set, and the rest of the data-frame and series became the train set. Since the train & test set for x data frame were still containing Label column that was duplicated in the y series, the Label column was dropped from the train & test set for x data-frame using commands xTrain[:,:-1] and xTest[:,:-1]. The CSE-CIC-IDS2018 dataset contained features that highly vary in magnitudes, units, and range; for example, some features counted the number of data packets, some counted the seconds of data flow, some had huge numbers, some had negative numbers. However, some machine learning models were highly sensitive to feature scaling due to the optimization techniques and calculating mechanisms. Therefore, the feature scaling played a significant role to keep the machine learning model training properly. MinMaxScaler was used to normalize the train & test set in the project. The preprocessing module provided standardization, normalization, transformation functions that converted datasets into suitable representations for machine learning models. *Pre-processing.MinMaxScaler()*, one of the data scalers provided by the preprocessing module, translated values of each feature of the train set and test set into the range between 0 and 1.

D. Feature Selection

The feature selection techniques extracted highly relevant features. The project applied two feature selection techniques, ANOVA F-test & RFE. ANOVA F-test & RFE have been imported with *f_classif* and *RFE* functions from sklearn library. A floating-point error was handled by ignoring zero and invalid floating-point operation division. SelectPercentile function, called selector, passed the *f_classif* function and defined the highest scoring percentage of features to 10%. The returned values of a selector, *x_f_train*, were the numbers of selected instances and features from the train set. Since RFE required a machine learning classifier to evaluate the feature importance, the decision tree classifier was created first and named with clf. *n_features_to_select* parameter, the number of features selected from the train set, was set to 5, corresponding to the number of features that could reach the best accuracy based on feature ranking from the RFECV function. *x_rfe_train* variable was the returned values of selected instances and features.

E. Building the Model

This stage was to build the decision tree classifier. Machine learning model was the key component of the project. Thanks to sklearn library, the decision tree classifier was directly called from the sklearn.tree module without complex and tedious coding process. The function calling decision tree classifier was *DecisionTreeClassifier()* and named *clf_all*. *Fit()* function fed the train set and corresponding labels into the machine learning model for training.

F. Prediction and Evaluation

This stage was to predict the network attacks using test data and evaluate the machine learning model along with the proposed feature selection method. The machine learning model was built and trained using train set, the model was ready to produce the final prediction based on the test set. *predict()* function provided by sklearn library was passed to

trained machine learning model, *clf_all*, and predicted the label value for all samples in the test set, xTest. The predicted output returned by *predict()* function was stored into the new variable called *Y_all_pred*. Various test scores were calculated with cross validation strategy. make_scorer module made scorers based on the performance metric. Multiple scoring functions, including accuracy, *precision_score*, *recall_score*, and *f1_score*, were imported to compute corresponding scores. Cross validation was employed to avoid the potential impact of small samples for some targets, for example, there were only 8 instances of SQL Injection in the test set. *cross_validate* function split the dataset into four smaller sets with same label distribution. Among four equal-sized sets, three sets were used to train the model and the remaining one was used to calculate the performance metrics. A customized function, average_score_on_cross_val_classification, was defined to evaluate the machine learning model using cross_validate function and return the absolute mean value for all four scores.

V. Results and Discussion

This section introduces the results of the study. Dataset presentation section was provided to describe the dimension of the CSE-CIC-IDS2018 dataset and sample view of the dataset. Feature construction section introduced the features that were constructed using the data provided in the original dataset. Feature selection section described the feature selection methods used in the project and listed the selected features. Model evaluation explained the performance metrics based on the different feature selection methods.

A. Feature Construction

Feature construction was completed by using CICFlowMeter, a feature extractor that extracted information from the bidirectional flows. CICFlowMeter provided functions to create time-related features from Pcap files of both forward and backward flows. In the original Pcap files, seven raw features presented the sequence of the data flows and all sorts of packet information. Seven raw features included FlowID, Timestamp, SourceIP, DestinationIP, SourcePort, DestinationPort, and Protocol. However, SourceIP and DestinationIP were removed due to the possibility of data leaks during the model training and the difficulty of feature encoding. Using the feature construction of CICFlowMeter extra 76 features were identified for various network threats like traffic rate, message size, duration among messages, TCP flags, header size and fragment length, preliminary window, active time and waiting time in forward and reversed streams, respectively. Some samples of derived features and descriptions were shown in Table IV.

B. Dataset Presentation

After loading the CSE-CIC-IDS2018 dataset, Executing df.shape function presented the dimension of the imported dataset. The output showed that the created data frame contained 5138535 instances and 80 columns, which included 79 features and one label column. The statistical summary of the CSE-CIC-IDS2018 dataset included the count of values, unique values, top values, and frequency of occurrence in each column, as shown in Fig. 3.

TABLE IV. Samples of Derived Features

Feature Name	Description
Flow Byte/s	Flow rate in bytes per second (bps)
Flow Pkts/s	Flow rate in packets per second
Flow IT Mean	Inter-packet interval mean time
Flow IT Max	The flow's longest possible interval between packets
Flow IT Min	The shortest possible interval between two packets in a flow
Fwd IT Min	The min amount of time between forward flow packets.
Fwd IT Max	The max amount of time between forward flow packets.
Fwd IT Mean	Packet forwarding averaging time
Fwd IT Total	Between-packet time in the forward flow
Bwd IT Min	Between-packet intervals in the reversed flow
Bwd IT Max	Between-packet intervals in a reversal flow
Bwd IT Mean	reverse flow packet-to-packet delay
Bwd IT Total	In reversed flow, the total amount of time between each packet.
FIN Pkts	The number of FIN packets in the stream
SYN Pkts	A flow's number of SYN packets
RST Pkts	There are a certain number of RST packets in the flow.
PSH Pkts	A flow's number of PSH packets
ACK Pkts	The flow's ACK packet count
URG Pkts	Quantity inflow of URG packets
CWR Pkts	How many packets of CWR are in the stream
ECE Pkts	Data packets per second (DPP)
Fwd Sgmt Size Avg	Dimensions of a typical forward flow PDU segment
Bwd Sgmt Size Avg	Dimensions of the PDU segments in the reversed flow
Fwd Byte Bk Avg	The forward flow of bytes has an average bulk.
Fwd Pkt Bk Avg	In the forward flow, the average number of packets.
Bwd Byte Bk Avg	In a reversed flow, the bulk of the average bytes
Bwd Pkt Bk Avg	In the reversed flow, the average number of packets per second.

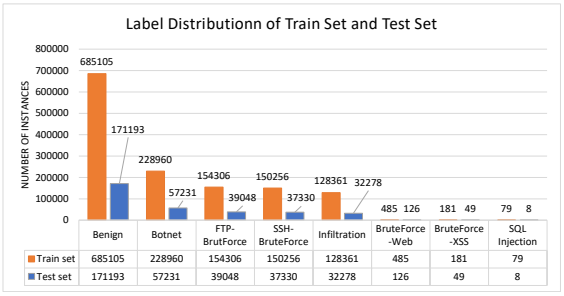


Fig. 3. Label Distributionn of Train Set and Test Set

In the data preprocessing and scaling process, timestamp information was removed from the dataset due to the insignificance for model training; all remaining data was converted to float datatype; all instances that contained Infinity and NaN values were dropped from the dataset; long decimal digits were reduced to one decimal digit for saving training time and memory; the attack labels were renamed with numbers as shown in Table V.

TABLE V. Label Names and Corresponding Numbers

Label	Number
Benign	0
Bot	1
FTP-BruteForce	2
SSH-Bruteforce	3
Infiltration	4
Brute Force -Web	5
Brute Force -XSS	6
SQL Injection	7

Then training and testing were created using the dataset. The training set has been created with 1,347,733 instances and 78 features, and the test set had 337,263 instances with

78 features. The train and test sets' label distribution shows a huge amount of benign traffic in the CSE-CIC- IDS2018 dataset. Compared with benign traffic, the number of attack traffic was relatively small. Lack of balance between different kinds of network attacks and benign traffic became the biggest weakness of the CSE-CIC-IDS2018 dataset.

C. Feature Selection

ANOVA F-test and Recursive attribute removal were used to choose features for this study. Using only the test results, the ANOVA F-test was used to determine whether each attribute connected to the label. An ML technique, such as a decision tree classifier in this instance, was employed as a wrapper selection approach to exclude features based on the significance of each feature to prediction results.

ANOVA F-test feature selection tested if each feature had any impact on classifying the attack categories. With one-way ANOVA, the p-value was computed to verify the likelihood that an attacker group could be accurately labelled based solely on the results of every feature. P-values larger than 0.05 indicated a stronger relationship with a specific attack category. All features were ranked in descending order based on the p-values. As this work decided to select the top 10% of the ranked features, the first eight features were selected as the feature subset. Table VI showed eight features selected by ANOVA F-test. The eight chosen features by ANOVA F-test were all constructed features, proving that the constructed features had a better statistical relationship with the attack categories than raw features.

TABLE VI. SELECTED FEATURES BY ANOV F-TEST

Label Index	Feature Name
16	Flow Pkts/s
37	Fwd Pkts/s
38	Bwd Pkts/s
46	RST Pkts
49	URG Pkts
51	ECE Pkts
66	Init Fwd Win Byts
69	Fwd Sgmt Size Min

Recursive Feature Elimination created the feature subset by evaluating the feature importance through the machine learning estimator. Training the decision tree predictor with the most attributes was necessary to determine each feature's importance. Due to the sample splitting mechanism, the decision tree algorithm introduced a built-in function, such as Gini Impurity, for calculating the feature importance in terms of the misclassification rate. The feature with lower Gini Impurity was preferred and significant because the misclassification rate of this feature was lower. Once this was done, the most insignificant feature was eliminated from the present extracted features. Five features were needed to complete the iterative training process and remove candidates. Table VII lists the top five features that were eliminated using Recursive Feature Elimination (RFE). The five features selected by Recursive Feature Elimination showed the low misclassification rate during the training process of the decision tree model.

D. Model Evaluation

This work calculated the confusion matrix's test scores with various experiments' cross-verification procedure. The ML model's positive sample classification accuracy was evaluated using the precision score. The capacity to correctly identify all positive samples was referred to as a recall. In order to calculate the F1 score, the weighted average of the precision and recall scores was taken into account.

TABLE VII. SELECTED FEATURES BY RECURSIVE FEATURE ELIMINATION

Label Index	Feature Name
0	Dst Port
26	Bwd IT Tot
28	Bwd IT Std
38	Bwd Pkts/s
69	Fwd Sgmt Size Min

This section discusses the classification report of the machine learning model in four comparison experiments. The classification report presented the precision, recall, and f1 score for each kind of network attack and benign traffic individually on the top: the macro average and weight average of test scores and overall accuracy on the bottom. There were 8 numbers from 0 to 7 on the top left of the report. Number 0 was the benign traffic, and numbers 1 to 7 represented seven network attacks, respectively.

Fig. 4 showed the performance metrics of the decision tree model that used 7 raw features to classify each kind of network attack and benign traffic with a cross-validation strategy.

	precision	recall	f1-score	support
0	0.83	0.92	0.87	861502
1	0.81	0.01	0.03	57221
2	0.00	0.00	0.00	38616
3	0.00	0.00	0.00	37614
4	0.03	0.04	0.03	32547
5	0.00	0.00	0.00	128
6	0.04	0.04	0.04	50
7	0.00	0.12	0.00	17
accuracy			0.77	1027695
macro avg	0.21	0.14	0.12	1027695
weighted avg	0.74	0.77	0.73	1027695

Fig. 4. Classification Report using 7 Raw Features

Fig. 5 showed the performance metrics of the decision tree model that used all raw features and constructed features to classify each kind of network attack and benign traffic with a cross-validation strategy.

	precision	recall	f1-score	support
0	0.94	0.76	0.84	855969
1	1.00	0.99	0.99	57231
2	1.00	1.00	1.00	39048
3	1.00	0.50	0.67	37330
4	0.05	0.28	0.08	32278
5	0.83	0.40	0.54	126
6	0.00	0.94	0.01	49
7	0.00	0.00	0.00	8
accuracy			0.76	1022039
macro avg	0.60	0.61	0.52	1022039
weighted avg	0.92	0.76	0.83	1022039

Fig. 5. The Classification Report using All Raw and Constructed Features

Fig. 6 showed the performance metrics of the decision tree model that used the ANOVA F-test to obtain the feature subset from all raw and constructed features and classified each kind

of network attack and benign traffic with a cross-validation strategy.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	855969
1	0.99	0.97	0.98	57231
2	1.00	1.00	1.00	39048
3	1.00	1.00	1.00	37330
4	0.35	0.03	0.06	32278
5	0.89	0.27	0.41	126
6	0.75	0.31	0.43	49
7	0.00	0.00	0.00	8
accuracy			0.97	1022039
macro avg	0.74	0.57	0.61	1022039
weighted avg	0.95	0.97	0.95	1022039

Fig. 6. The Classification Report using ANOVA F-Test

Fig. 7 showed the performance metrics of the decision tree model that used Recursive Feature Elimination to obtain the feature subset from all raw and constructed features and classified each kind of network attack and benign traffic with cross-validation strategy.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	855969
1	1.00	1.00	1.00	57231
2	1.00	1.00	1.00	39048
3	1.00	1.00	1.00	37330
4	0.24	0.10	0.14	32278
5	0.98	0.33	0.49	126
6	0.96	0.47	0.63	49
7	0.07	0.12	0.09	8
accuracy			0.96	1022039
macro avg	0.78	0.63	0.67	1022039
weighted avg	0.95	0.96	0.95	1022039

Fig. 7. The Classification Report using RFE

In Fig. 8, the overall test scores of the decision tree model using constructed features and feature selection techniques were much better than using raw features. Especially combing feature selection and constructed features reached more than 95% on all test scores. But using raw features only got lower than 80% for all test scores.

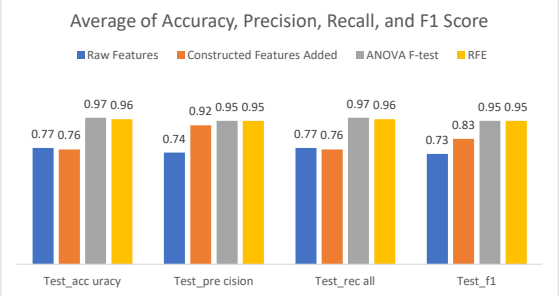


Fig. 8. Weighted Average of Accuracy, Precision, Recall, and F1 Score for Four Experiments

This part provided a comparison of the precision, recall, and f1 scores for predicting different kinds of network attacks in four experiments. By looking at Fig. 9, Fig. 10, and Fig. 11, the decision tree model that used ANOVA F-test and Recursive Feature Elimination to obtain the feature subset presented good performance in detecting botnet attack, FTP-BruteFrce, and SSH-BruteForce attack. The precision, recall, and f1 score of detecting botnet attack, FTP-BruteFrce attack, and SSH-BruteForce attack when using feature selection techniques and constructed features reached more than 97%, which were way better than the test scores using raw features.

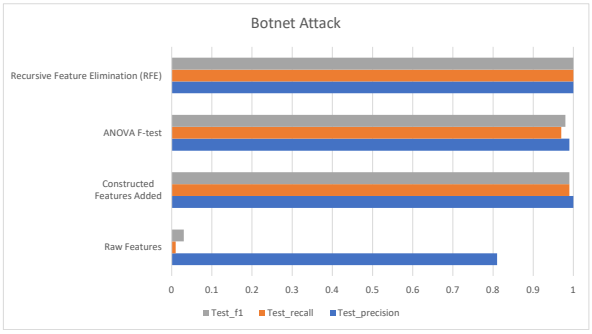


Fig. 9. Test Scores of Four Experiments on Botnet Attack

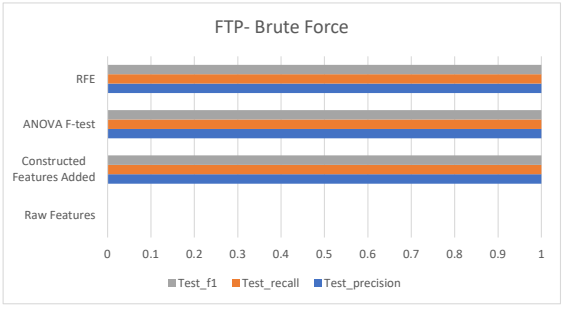


Fig. 10. Test Scores of Four Experiments on FTP-BruteForce

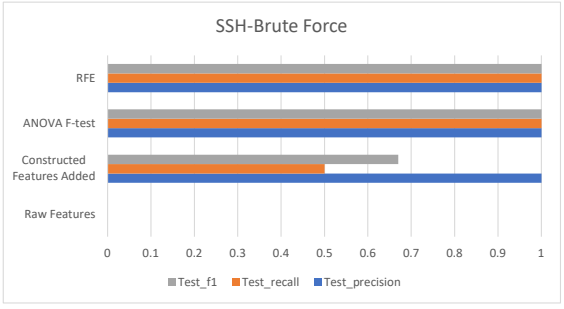


Fig. 11. Test Scores of Four Experiments on SSH-BruteForce

However, when detecting infiltration attack, web attack, and SQL injection, the precision, recall, and f1 score did not increase dramatically with feature selection and constructed

features shown in Fig. 12, Fig. 13, and Fig. 14. But still, the test scores using feature selection and constructed features were slightly better than using raw features.

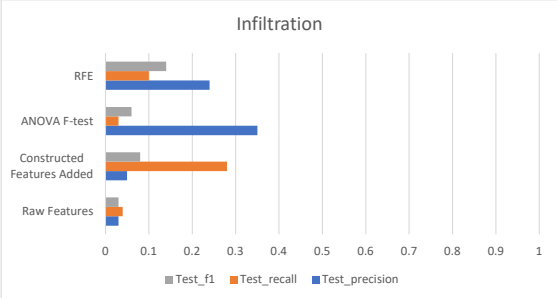


Fig. 12. Test Scores of Four Experiments on Infiltration Attack

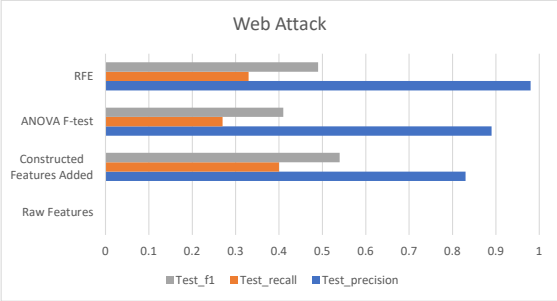


Fig. 13. Test Scores of Four Experiments on Web Attack

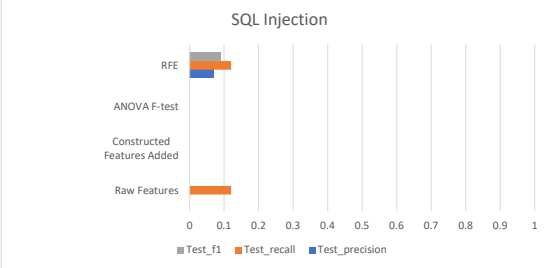


Fig. 14. Test Scores of Four Experiments on SQL Injection

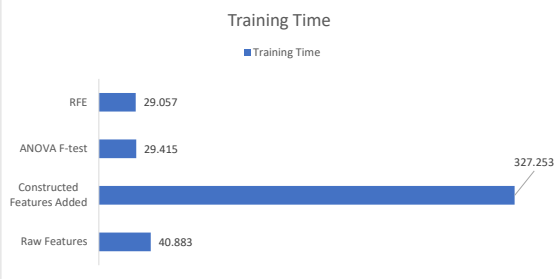


Fig. 15. Training Time of Four Experiments

In Fig. 15, adding constructed features to raw features dramatically increased the training time and took over 327 seconds to train the decision tree model. More data was added to the dataset, requiring the decision tree model to spend more time processing them.

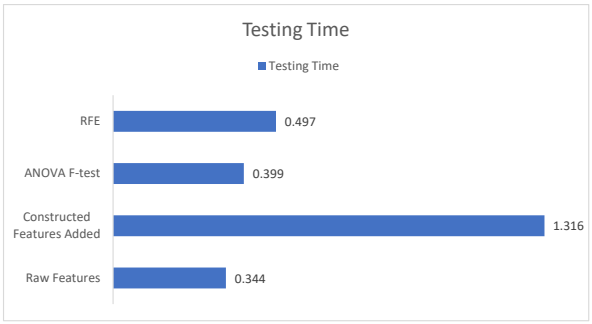


Fig. 16. Testing Time of Four Experiments

However, after using ANOVA F-test and Recursive Feature Elimination, the training time was reduced to around 29 seconds, even lower than the time used for training with raw features. In Fig. 16, the testing time presented a similar situation as the training time shown in Fig. 15. The combination of constructed features and raw features made the decision tree model spend much more time to produce the prediction result than only using raw features or employing two feature selection techniques. However, the testing time of using feature selection techniques was slightly longer than using raw features.

VI. Conclusion

Feature selection and derived features were combined in this work to improve ML model performance in NIDS. A characteristic attack feature was constructed using CICFlowMeter, and a feature subset was created using ANOVA F-test and Recursive Feature Elimination to achieve the goal. Using the CSE-CIC-IDS2018 dataset from the Canadian Institute for Cybersecurity and Communications Security Establishment, the project used a decision tree machine learning model to detect network threats. Python was used on Jupyter Notebook to create the machine learning model and the testing environment. Evidence suggests that the combination of feature selection techniques and derived features can improve prediction precision accuracy recall F1 score and the decision tree model's prediction accuracy and precision. The CSE-CIC-IDS2018 dataset used in the project was the most recent benchmark intrusion detection dataset. The dataset provided a large number of samples for various kinds of popular network attacks in the Internet, including Brute Force, Web attack, Infiltration, and Botnet attack. However, the only weakness of the CSE-CIC-IDS2018 dataset was that the sample distribution of network attacks was not quite balanced, which may influence the machine learning model's performance to some degree. In order to reduce the impact of imbalanced samples, the project employed the cross-validation technique to split the dataset into multiple folds that contained the same distribution of samples from different classes.

More samples for different kinds of network attacks will be collected to enrich and balance the current CSE-CIC-IDS2018 dataset in future work. Since this work only tested ANOVA F-test and RFE, in the future additional feature selection techniques will be used to investigate new combinations for better network intrusion detection predictions.

REFERENCES

[1] Y. Aleksieva, H. Valchanov, and V. Aleksieva, "An approach for host based botnet detection system," 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), 2019.

[2] Khraisat, A., Gondal, I., Vamplew, P. et al. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecur 2, 20 (2019). https://doi.org/10.1186/s42400-019-0038-7

[3] Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T. et al. Intrusion detection model using machine learning algorithm on Big Data environment. J Big Data 5, 34 (2018). https://doi.org/10.1186/s40537-018-0145-4

[4] Ebrahimi, H., Majidzadeh, K., Soleimanian Gharehchopogh, F. (2022). Integration of deep learning model and feature selection for multi-label classification. International Journal of Nonlinear Analysis and Applications, 13(1), 2871-2883. doi: 10.22075/ijnaa.2021.25379.2998

[5] Osanaiye, Opeyemi & Choo, Kim-Kwang Raymond & Dlodlo, Mqhele E.. (2016). Analysing Feature Selection and Classification Techniques for DDoS Detection in Cloud.

[6] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," Machine Learning, vol. 101, no. 1-3, pp. 59–84, Apr. 2014.

[7] S.-W. Lin, K.-C. Ying, C.-Y. Lee, and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," Applied Soft Computing, vol. 12, no. 10, pp. 3285–3290, 2012.

[8] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, 2018.

[9] Chen, RC., Dewi, C., Huang, SW. et al. Selecting critical features for data classification based on machine learning methods. J Big Data 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4

[10] Zhang, Jian, Qidi Liang, Rui Jiang, and Xi Li. 2019. "A Feature Analysis Based Identifying Scheme Using GBDT for DDoS with Multiple Attack Vectors" Applied Sciences 9, no. 21: 4633. https://doi.org/10.3390/app9214633

[11] J.-H. Woo, J.-Y. Song, and Y.-J. Choi, "Performance Enhancement of Deep Neural Network Using Feature Selection and Preprocessing for Intrusion Detection," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019.

[12] Y.-L. Wan, J.-C. Chang, R.-J. Chen, and S.-J. Wang, "Feature-Selection-Based Ransomware Detection with Machine Learning of Data Analysis," 2018 3rd International Conference on Computer and Communication Systems (ICCCS), 2018.

[13] P.-S. Tang, X.-L. Tang, Z.-Y. Tao, and J.-P. Li, "Research on feature selection algorithm based on mutual information and genetic algorithm," 2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing (ICCWAMTIP), 2014.

[14] G. Manikandan, E. Susi, and S. Abirami, "Feature Selection on High Dimensional Data Using Wrapper Based Subset Selection," 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), 2017.

[15] M. S. Kumar, J. Ben-Othman, K. Srinivasagan, and G. U. Krishnan, "Artificial Intelligence Managed Network Defense System against Port Scanning Outbreaks," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019.

[16] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, 'Building an efficient intrusion detection system based on feature selection and ensemble classifier', Comput. Networks, vol. 174, 2020.

[17] AHUJA, RAVINDER, and SC SHARMA. "Exploiting Machine Learning and Feature Selection Algorithms to Predict Instructor Performance in Higher Education." Journal of Information Science & Engineering 37.5 (2021).

[18] S. Cateni, V. Colla, and M. Vannucci, "A Hybrid Feature Selection Method for Classification Purposes," 2014 European Modelling Symposium, 2014.

[19] G. Smith, "Step away from stepwise," Journal of Big Data, vol. 5, no. 1, 2018.

[20] B. Sahu, S. Dehuri, and A. Jagadev, "A Study on the Relevance of Feature Selection Methods in Microarray Data," The Open Bioinformatics Journal, vol. 11, no. 1, pp. 117–139, 2018.

[21] L. Ladha and T. Deepa, "Feature selection methods and algorithms," International journal on computer science and engineering, 3(5), 1787-1797, 2011.

[22] J. Peltonen, "Lecture 2: Feature selection," Dimensionality Reduction and Visualization, 2014. [PowerPoint slides]. [Accessed: 28-Nov-2019].

[23] Q. Zhou and D. Pezaros, "Evaluation of Machine Learning Classifiers for Zero-Day Intrusion Detection–An Analysis on CIC-AWS-2018 dataset," arXiv preprint arXiv:1905.03685, 2019.

[24] I.-V. Onut and A. Ghorbani, "Toward A Feature Classification Scheme For Network Intrusion Detection," 4th Annual Communication Networks and Services Research Conference (CNSR06), 2007.

[25] M. M. Najafabadi, T. M. Khoshgoftaar, C. Kemp, N. Seliya, and R. Zuech, "Machine Learning for Detecting Brute Force Attacks at the Network Level," 2014 IEEE International Conference on Bioinformatics and Bioengineering, 2014.

[26] K. Wang, C.-Y. Huang, S.-J. Lin, and Y.-D. Lin, "A fuzzy pattern-based filtering algorithm for botnet detection," Computer Networks, vol. 55, no. 15, pp. 3275–3286, 2011.

[27] B. Setiawan, S. Djanali, and T. Ahmad, 'Increasing accuracy and completeness of intrusion detection model using fusion of normalization, feature selection method and support vector machine', Int. J. Intell. Eng. Syst., vol. 12, no. 4, pp. 378–389, 2019.

[28] Q. Liao, H. Li, S. Kang, and C. Liu, "Feature extraction and construction of application layer DDoS attack based on user behavior," Proceedings of the 33rd Chinese Control Conference, 2014.

[29] Xu, X. (2006). Adaptive Intrusion Detection Based on Machine Learning : Feature Extraction , Classifier Construction and Sequential Pattern Prediction.

[30] M. Ring, D. Landes, and A. Hotho, "Detection of slow port scans in flow-based network traffic," Plos One, vol. 13, no. 9, 2018.

[31] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, 1(2), 111-117, 2006.

Path Optimization for Mobile Robots using Genetic Algorithms

Fernando Martínez Santa, Fredy H. Martínez Sarmiento, Holman Montiel Ariza
Universidad Distrital
Francisco José de Caldas
Bogotá, Colombia

***Abstract*—This article proposes a path planning strategy for mobile robots based on image processing, the visibility graphs technique, and genetic algorithms as searching/optimization tool. This proposal pretends to improve the overall execution time of the path planning strategy against other ones that use visibility graphs with other searching algorithms. The global algorithm starts from a binary image of the robot environment, where the obstacles are represented in white over a black background. After that four *keypoints* are calculated for each obstacle by applying some image processing algorithms and geometric measurements. Based on the obtained *keypoints*, a visibility graph is generated, connecting all of these along with the starting point and the ending point, as well as avoiding collisions with the obstacles taking into account a safety distance calculated by means of using an image dilation operation. Finally, a genetic algorithm is used to optimize a valid path from the start to the end passing through the navigation network created by the visibility graph. This implementation was developed using *Python* programming language and some modules for working with image processing ang genetic algorithms. After several tests, the proposed strategy shows execution times similar to other tested algorithms, which validates its use on applications with a limited number of obstacles presented in the environment and low-medium resolution images.**

Keywords*—Optimization; path planning; genetic algorithms; visibility graphs

I. Introduction

Today more than ever, robotics is part of the daily life in most of the world specially in big cities where the automation and smart stuff is everywhere. Mobile robots area has been widely researched not only in its mechanical design and locomotion type but in its motion planning [1], [2] in applications such as movement in indoor environments [3], obstacle avoidance [4], navigation in complex mazes [5], path planning [6], [7], and some times using open source robotics software [8]. Researching areas like UAVs (Unmanned Aerial Vehicles) and self-driving or autonomous vehicles have maintained the interest on one of the most important issues for mobile robots, the path planning. In this area, one of the most used algorithm has been the visibility graphs [9] supported by image processing algorithms [10]. These visibility graphs generates a high dense network of possible paths through the some navigation keypoints obtained from the obstacles image that is related with the navigation scene to solve by the mobile robot. This path network includes also both the starting point and the ending point, then a valid and short path has to be found from the start to the end passing through some segments of the connection network, for that reason it is necessary to apply a decision or optimization algorithm [11] to choose the best path inside on that network. A lot of different optimization algorithms has been used for the path planning issue such as ant colony optimization [12], [13], particle swarm for mobile robots [14], [15], [16], [17], chaotic particle swarm [18] particle swarm for manipulators [19], brain storm optimization [20], Fuzzy-Wind Driven algorithm [21], rapidly-exploring trees [22], gray wolf algorithm [23] among others.

The Genetic Algorithms *GAs* are searching and optimization methods based on the natural selection process and the genetic operations involved in it, these ones have been used for solving problems in a lot of different engineering areas including robotics and of course path planning [24], [25], [26] and other variations such as the Taxi carpooling algorithm [27].

Therefore, this research aims to propose a path planning strategy combining both the visibility graphs method using image processing algorithms and genetic algorithms to optimize and obtain the best path, improving the overall execution time respect other related works. All of this strategy is proposed to be solved by means of using free software in this case all of the algorithms will be implemented on *Python* language using modules such as: *scikit-image* and *geneticalgorithm2*.

The paper is organized as follows: Section 2 presents the methodology proposed to find the optimal path for mobile robots using Genetic Algorithms (GAs), describing all processes to identify the obstacles image by capturing the image of navigating environment; going through the *keypoints* obtaining, then generating visibility graph, and subsequently selecting path optimization using GAs. Section 3 presents the results of implementing the path planning strategy in Python 3 language and testing in different navigation environments to evaluate overall execution time. Finally, Section 4 presents the conclusions about this research's main ideas, including possible future jobs.

II. Methodology

The path planning strategy for mobile robots proposed in this article is based on image capture of the navigating environment in which the robot is involved [28], [29], where the obstacles are perfectly differentiated from the void room. First step is the calculation of some *keypoints* for each obstacle in the scene in order to reduce the amount of information to work with by the planning algorithm. After that, all the possible paths from the start to the end are calculated, all of them passing through the previously detected *keypoints*, producing a

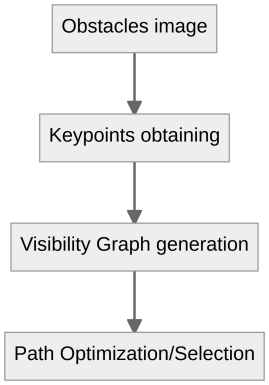


Fig. 1. General Algorithm Pipeline.

highly dense path network. Finally, the optimization algorithm [30] is applied to the path network in order to obtain the shortest path. The complete process can be sumarized through the flow chart shown in Fig. 1, where it starts from a binary image of the environment (a black background and white obstacles), then some *keypoints* by each obstacle are calculated, after that a visibility graph is generated and finally only one path is selected. In the next sections, each step of the process is explained in detail.

A. Obstacles Image

The proposed strategy starts with a Black & White image of the scene with the obstacles [31], this binary image has a black background and the obstacles are represented in white as the example shown in Fig. 2. This image can be obtained from a camera located at the top of the robot environment and after turned into binary by means of applying an image threshold operation.

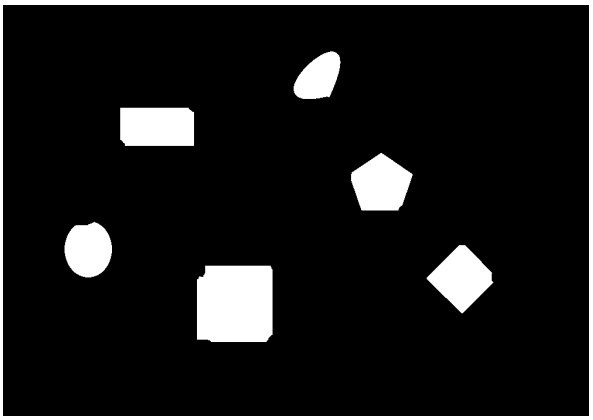


Fig. 2. Starting Scene Binary Image.

B. Keypoints Obtaining

The *keypoints* obtaining step is supported by some digital image algorithms, first the binary image of the obstacles is dilated in order to expand the obstacles border, applying the correspondent image morphology operation. After, the main two axis and the centroid of each dilated obstacle are calculated, this is done labeling and measuring each separated region in the binary image (obstacles). Finally, four *keypoints* per obstacle are calculated.

1) Image Dilation: In this proposal, the navigation *keypoints* are based on the obstacles borders, but this takes into account a safe distance between these ones and the robot [32]. That distance is calculated from to the maximum radius of the robot according to the eq. 1 and correspons to the dilation radius r_d.

$$r_d = \lfloor r_m + \Delta r \rfloor \tag{1}$$

Where r_m is the maximum radius of the robot and Δr is a radius tolerance defined at 10% in this case. Finally, r_d has to be an integer and it is represented in pixel units.

Once the dilation radius r_d is obtained, the morphology dilation operation is performed on the obstacles image by means of applying a $2D$ convolution between the original binary image and a square shape as wide as r_d. The result is shown in Fig. 3, where the obstacles are the same as the binary image (see Fig. 2), but their area is expanded because of the dilation operation. This operation allows assuming obstacles with major areas to avoid future collisions due to the maximum radius of the mobile robot.

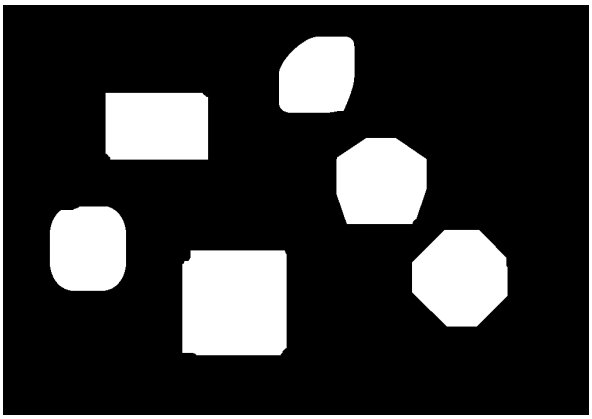


Fig. 3. Dilated Obstacles Image.

2) Obstacle keypoints Computing: After dilating the obstacles image, each obstacle is labeling and measured in order to find its centroid and its two main axis, from this data a Δx and a Δy are calculated for each axis according to eq. 2.

$$\forall i \in I \quad : \begin{cases} \Delta x_i = \cos(\theta) \cdot \left(\frac{l_i}{2}\right) \\ \\ \Delta y_i = \sin(\theta) \cdot \left(\frac{l_i}{2}\right) \end{cases} \tag{2}$$

Where l_i is the length of each main axis i of the obstacles axis set I and θ the orientation of the major axis detected. Finally from these deltas, the *keypoints* are calculated according to the eq. 3.

$$\forall i \in I,\ \forall j \in J \quad : \begin{cases} x_{ij} = x_0 \pm \Delta x_i \\ \\ y_{ij} = y_0 \pm \Delta y_i \end{cases} \tag{3}$$

Where j is each of the calculated points set J for each main axis set I and (x_0, y_0) is the centroid of each obstacle. The cardinality of the sets I and J is $|I| = |J| = 2$, so for each obstacle two main axis are calculated, and for each of them two points are generated, for a total of 4 generated *keypoints* by obstacle, as shown in Fig. 4.

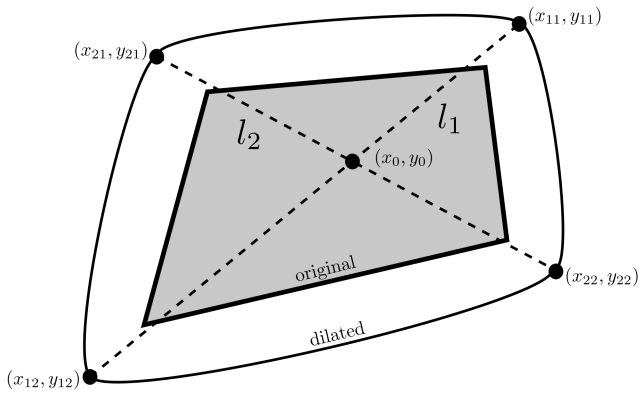


Fig. 4. Obstacle *keypoints* Computing.

The Fig. 4 shows schematically the dilation process. Having an diamond-shaped obstacle (shown in gray) in the original image, the external rounded shape (in white) represents the same obstacle after the image dilation process. The main axis are represented by dashed lines, (x_0, y_0) is the centroid and the (x_{ij}, y_{ij}) are the obtained *keypoints*. The obtained *keypoints* along with the original obstacles are shown in Fig. 5, where the starting point is at the lower-left corner and the ending point is at the right border.

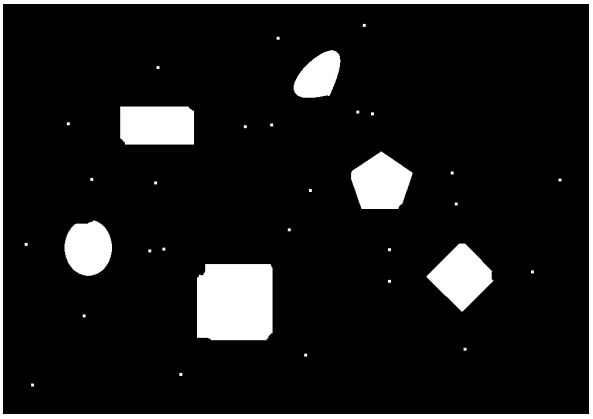


Fig. 5. Navigable *keypoints* Image.

C. Visibility Graph Generation

After generating all of the navigating *keypoints* including the starter and ending points, the visibility graph [33] is generated (see Fig. 6) connecting all of these points (drawing lines on the image) and after avoiding the crossing with the obstacles as shown in Fig. 7.

The collision avoidance between these lines and the obstacles is calculated by means of applying binary image operations, specifically a binary *XOR* operation (exclusive disjunction) between the dilated obstacles image and a copy of the same image with the specific line drawn in *black*. If there is a collision, a black segment line will appear over an obstacle, that means that the two images will be different. Both images have to be totally equals (pixel by pixel) for validating the line, so each resultant pixel has to accomplish the eq. 4,

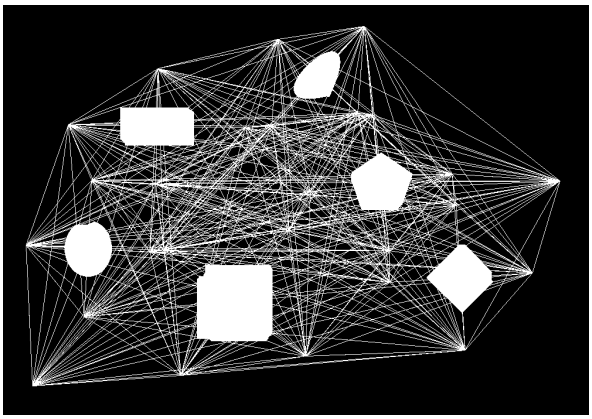


Fig. 6. Visibility Graph.

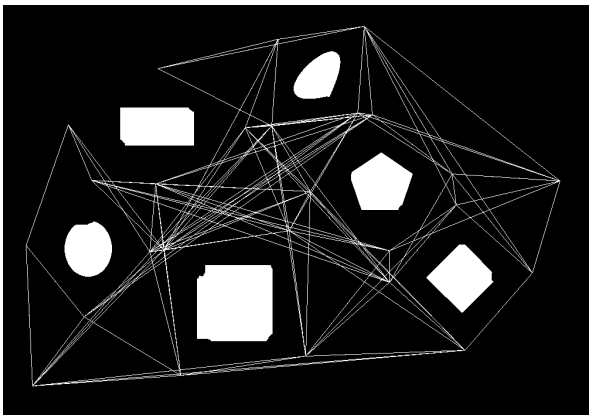


Fig. 7. Visibility Graph (Avoiding Collisions).

$$\forall i \in I, \forall j \in J \quad : A_{ij} \nleftrightarrow B_{ij} = \textit{False} \tag{4}$$

Where I and J for this equation, are the sets of all valid indices in both dimensions of the dilated obstacles image A and the image B which has the dilated obstacles plus the drawn line.

D. Path Optimization-Selection

For selecting the shortest possible path in the generated visibility graph, it is possible to use a lot of different selection or optimization algorithms such as the A^* algorithm [34]. In this proposal, Genetic Algorithms (GAs) are used as optimization tool [35] in order to find the shortest (and then the most efficient) path in the dense navigating network generated by the visibility graph.

1) Target Function: Once defined the complete set of *keypoints* P including the start p_s and the end p_e, it is necessary to define the *target function* to optimize $f(P)$, this depends on the cumulative distance of each segment $\overline{p_i p_{(i+1)}}$ from p_s to p_e, that accomplishes with the eq. 4. For the target function definition, it is necessary to define the solution set X as shown in eq. 5, where each x represents an index for reading the *keypoints* set P, so the set X detemine the order of a subset $P_X \subseteq P$ which is a possible solution (a short path).

$$X = \{x_0, x_1, x_2 \dots x_n\} \tag{5}$$

The number of indices of the solution set X corresponds to the number of objects in the scene, plus the inital and ending points thus $n =$ *n-obstacles* $+ 2$. The number of elements of the set X is less or equal of the number of elements of the subset P_X, so $|P_X| \leq |X|$, This occurs because a P_{x_i} different from Px_n equals the ending point P_e, that is meant the path reaches the final in less steps than the maximum allowed n, so it is necessary to determine the real number of steps m. This last is possible applying the eq. 6.

$$\forall i \in \{0, 1, 2 \dots n\} \ : \ p_{x_i} = p_e \implies m = i \tag{6}$$

Once the steps m have been computed, the optimization target function is defined from the solution set X as shown in eq. 7.

$$f(X) = \sum_{i=0}^{m} |\overrightarrow{p_{x_i} p_{(x_i+1)}}| \tag{7}$$

Where $|\overrightarrow{p_{x_i} p_{(x_i+1)}}|$ is the distance of a segment between two sequential *keypoints*. This target function is subject to the first element of P_k were the starting point p_s, then it is generating the optimization restriction shown in eq. 8.

$$p_{x_0} = p_s \tag{8}$$

Additionally, as were described in the visibility graph section, the target function also has an obstacle collision restriction which can be implemented applying the eq. 4

2) Genetic Algorithm Implementation: The genetic algorithm proposed in this article for optimizing the visibility graph is setup as follows: there is no limit for the maximum number of iterations, the number of iterations without any improvement in the target function (fitness function) is set in 20, the crossover and mutation type era defined as uniform, the selection type is roulette, a 100 individuals population is defined, the rest of parameters are shown in Table I.

The *fitness function* or target function to minimize by the genetic algorithm is implemented according to the eq. 7 and specifying the restrictions (see eq. 4 and eq. 8), generation a penalty when one of them is not accomplished, that is meant $f(P)$ is carried to a maximum value $f(P)_{max}$ which corresponds to the total perimeter of the image A, then $f(P)_{max} = A_{x_{max}} * 2 + A_{y_{max}} * 2$.

TABLE I. GENETIC ALGORITHM PARAMETERS

Parameter	value
maximum iteration number	None
population size	100
mutation probability	0.1
elitism ratio	0.01
crossover probability	0.5
parents portion	0.3
crossover type	uniform
mutation type	uniform by center
selection type	roulette
max. iteration without improvement	20
dimension	same number of obstacles
variable type	integer
function timeout	10s

On the other hand the genome is built from the set X, taking into account that $\forall i \in \{0, 1, 2 \dots n\} \ : \ x_i \in \mathbb{E}$, the genome is simply generated in order as shown in Table II, being each gene an integer variable. No other variable different to the x_i is necessary to be appended to the genome.

TABLE II. GENOME ORDER.

gen_0	gen_1	gen_2	gen_3	$\dots$	gen_{n-1}
x_0	x_1	x_2	x_3	$\dots$	x_{n-1}

III. RESULTS

All of the proposed path planning strategy was implemented in *Python 3* language on a simple office laptop (whose features are described in Table III) running a *GNU/Linux* distribution. In total 10 tests were done over different navigation environments, all of them with the 6 obstacles as shown in most of figures. In average only the genetic algorithm execution time was in the interval of $(65, 94)s$ with an average value of $81.3s$ which is comparable with other previously tested algorithms such as A^* which showed over the same conditions an average execution time of $83s$.

TABLE III. TESTING PC FEATURES.

CPU	AMD Athlon Gold 3150U @ 2.400GHz
cores	2 hardware (4 subprocess)
GPU	AMD ATI 03:00.0 Picasso
RAM	12 GB
main drive	SSD
OS	Ubuntu 20.04.3 LTS x86_64

The genetic algorithm took in average 51.5 generations (iterations) to reach an average $f(X) = 1115$, the Fig. 8 shows a plot of one of the tests done, where the best target function in each generation is plotted. This specific example, reaches the convergence value in 44 generations. For the specific algorithm setup used, never this one reaches the maximum number of iterations (200), always this stops by reaching the maximum number of iterations without improvement the fitness function (20), that is meant the algorithm rapidly reaches a minimum but this is not necesary the global (see Fig. 9).

In order to find the global minimum a new test was done, this time with no limits of number of generations and generations without improvement. As a result this last test gave an execution time of $21min\ 33s$, almost 16 times more than the original tests, and this one reaches a $f(X) = 1027.7$

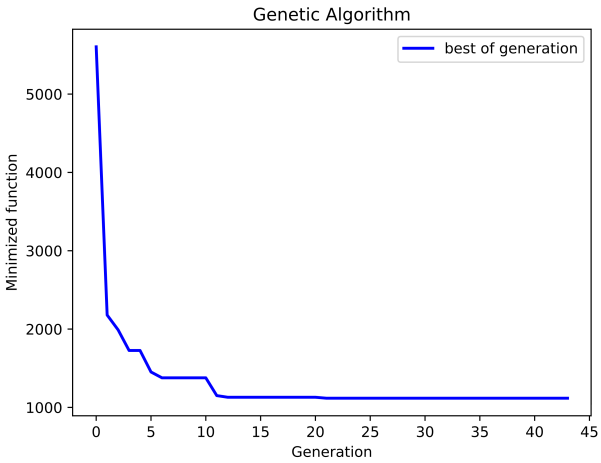


Fig. 8. Target Function Minimizing with a Limit of 20 Generations without Improvements.

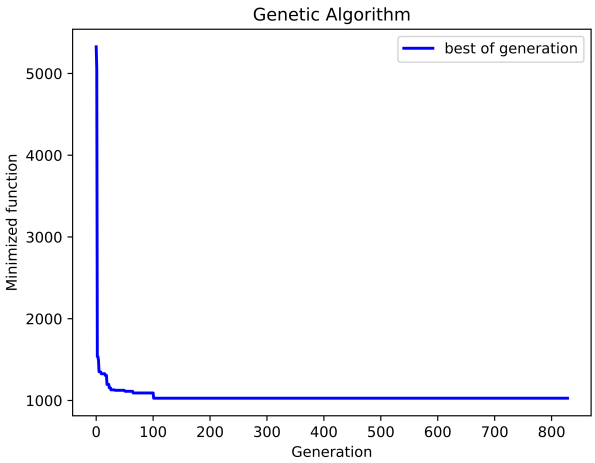


Fig. 10. Target Function Minimizing without a Limit of Generations.

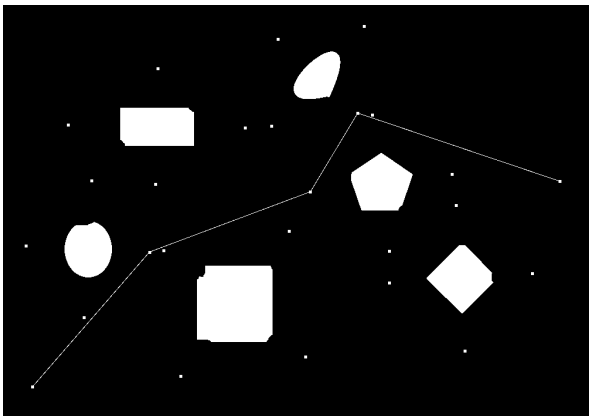


Fig. 9. Obtained Final Path Example.

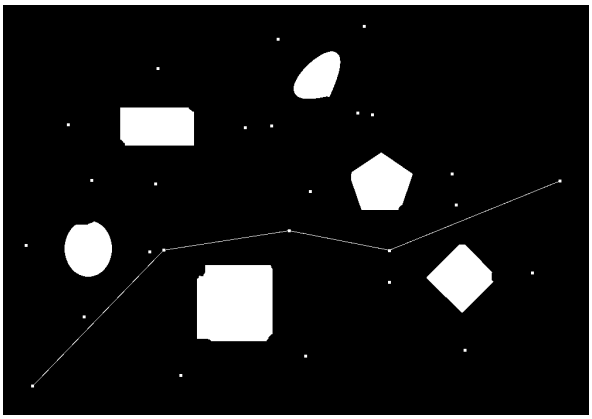


Fig. 11. Obtained Final Path, the Second Example.

which represents an improvement of 7.83%, obtained after 829 generations. The minimizing plot of this second test is shown in Fig. 10, where it is possible to graphically realize that the target function value has practically no changes from the generation number 110. The final path generated by this last test is shown in Fig. 11.

On the other hand, the image processing operations in charge of generating the collision restriction, spend around of the 43% of the total execution time of the genetic algorithm, mainly due to their use of pixel-to-pixel image comparisons which have a high computational cost. This execution time, can be exponentially increased by a linear increasing in the image dimensions.

IV. Conclusion

The proposed strategy of path planning based on visibility graphs and genetic algorithms, gave as a result execution times similar to other previously tested algorithms, in cases of simple environments such as the ones with a low amount of obstacles. According to the optimization parameters and environment images used, the genetic algorithm finds a valid solution with workable execution times. If the resolution of the environment input image or the number of obstacles increases, this proposed strategy will reach high execution times which will make it difficult to apply it on a real time robotics navigation task. This last could happen also is the application needs that the genetic algorithm finds the global minimum.

This proposal implements a safety distance between the obstacles and the mobile robot, this distance is based on an image dilation operation, this technique can limit the possible paths in the visibility graph as shown in Fig. 7 where a connection line is missing for the obstacle in the upper-left corner, as well as obstacles with concave shape could generate issues due to the methodology used for calculating the *keypoints*.

As future work, it is proposed to generate and automatic image resizing (without information losing) in order to reduce its dimensions and therefore the computing time involved in the image operations within the target function to be solved by the genetic algorithm. Another future work proposal is to develop a path planning strategy that uses directly the genetic algorithm over all the navigating free space of the image, so that the solution set will be not an index set but a set of (x, y)

direct points on the image. This last proposal, could take better advantage of the searching features of the genetic algorithms.

ACKNOWLEDGMENT

This work was supported by Universidad Distrital Francisco José de Caldas, specifically by the Technology Faculty. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank the ARMOS research group for the simulations and tests done.

REFERENCES

[1] B. Patle, A. Pandey, D. Parhi, A. Jagadeesh *et al.*, "A review: On path planning strategies for navigation of mobile robot," *Defence Technology*, vol. 15, no. 4, pp. 582–606, 2019.

[2] R. Manzoor and N. Kumar, "Mobile robot path planning approaches: Recent developments," in *Innovations in Information and Communication Technologies (IICT-2020)*, P. K. Singh, Z. Polkowski, S. Tanwar, S. K. Pandey, G. Matei, and D. Pirvu, Eds. Cham: Springer International Publishing, 2021, pp. 301–308.

[3] A. V. Rendón, "Evaluación de estrategia de navegación autónoma basada en comportamiento reactivo para plataformas robóticas móviles," *Tekhnê*, vol. 12, no. 2, pp. 75–82, 2015.

[4] N. Adzhar, Y. Yusof, and M. A. Ahmad, "A Review on Autonomous Mobile Robot Path Planning Algorithms," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 3, pp. 236–240, 2020.

[5] D. A. R. Ramos, B. A. B. Ruiz, and D. A. M. Osuna, "Desarrollo de control e implementación de robot móvil con la capacidad de solucionar laberintos complejos," *# ashtag*, no. 9, pp. 21–33, 2016.

[6] S. A. Zanlongo, L. Bobadilla, and Y. T. Tan, "Path-planning of miniature rovers for inspection of the hanford high-level waste double shell tanks," in *Florida Conference on Recent Advances in Robotics (FCRAR)*, 2017.

[7] H.-y. Zhang, W.-m. Lin, and A.-x. Chen, "Path planning for the mobile robot: A review," *Symmetry*, vol. 10, no. 10, p. 450, 2018.

[8] F. H. M. Sarmiento and D. A. G. Ramírez, "Openrrarch: una arquitectura abierta, robusta y confiable para el control de robots autónomos," *tecnura*, vol. 21, no. 51, pp. 96–104, 2017.

[9] L. Blasi, E. D'Amato, M. Mattei, and I. Notaro, "Path planning and real-time collision avoidance based on the essential visibility graph," *Applied Sciences*, vol. 10, no. 16, p. 5613, 2020.

[10] N. Roy, R. Chattopadhay, A. Mukherjee, and A. Bhuiya, "Implementation of image processing and reinforcement learning in path planning of mobile robots," *International Journal of Engineering Science*, vol. 15211, 2017.

[11] M. N. Zafar and J. Mohanta, "Methodology for path planning and optimization of mobile robots: A review," *Procedia computer science*, vol. 133, pp. 141–152, 2018.

[12] M. Brand, M. Masuda, N. Wehner, and X.-H. Yu, "Ant colony optimization algorithm for robot path planning," in *2010 international conference on computer design and applications*, vol. 3. IEEE, 2010, pp. V3–436.

[13] K. Akka and F. Khaber, "Mobile robot path planning using an improved ant colony optimization," *International Journal of Advanced Robotic Systems*, vol. 15, no. 3, p. 1729881418774673, 2018.

[14] K. Su, Y. Wang, and X. Hu, "Robot path planning based on random coding particle swarm optimization," *International journal of advanced computer science and applications*, vol. 6, no. 4, pp. 58–64, 2015.

[15] H. Mo and L. Xu, "Research of biogeography particle swarm optimization for robot path planning," *Neurocomputing*, vol. 148, pp. 91–99, 2015.

[16] T. T. Mac, C. Copot, D. T. Tran, and R. De Keyser, "A hierarchical global path planning approach for mobile robots based on multi-objective particle swarm optimization," *Applied Soft Computing*, vol. 59, pp. 68–76, 2017.

[17] X. Li, D. Wu, J. He, M. Bashir, and M. Liping, "An improved method of particle swarm optimization for path planning of mobile robot," *Journal of Control Science and Engineering*, vol. 2020, 2020.

[18] A. Tharwat, M. Elhoseny, A. E. Hassanien, T. Gabel, and A. Kumar, "Intelligent bézier curve-based path planning model using chaotic particle swarm optimization algorithm," *Cluster Computing*, vol. 22, no. 2, pp. 4745–4766, 2019.

[19] A. Machmudah, S. Parman, and M. Baharom, "Continuous path planning of kinematically redundant manipulator using particle swarm optimization," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 3, pp. 207–217, 2018.

[20] E. Dolicanin, I. Fetahovic, E. Tuba, R. Capor-Hrosik, and M. Tuba, "Unmanned combat aerial vehicle path planning by brain storm optimization algorithm," *Studies in Informatics and Control*, vol. 27, no. 1, pp. 15–24, 2018.

[21] A. Pandey and D. R. Parhi, "Optimum path planning of mobile robot in unknown static and dynamic environments using fuzzy-wind driven optimization algorithm," *Defence Technology*, vol. 13, no. 1, pp. 47–58, 2017.

[22] K. Qian, Y. Liu, L. Tian, and J. Bao, "Robot path planning optimization method based on heuristic multi-directional rapidly-exploring tree," *Computers & Electrical Engineering*, vol. 85, p. 106688, 2020.

[23] M. Radmanesh, M. Kumar, and M. Sarim, "Grey wolf optimization based sense and avoid algorithm in a bayesian framework for multiple uav path planning in an uncertain environment," *Aerospace Science and Technology*, vol. 77, pp. 168–179, 2018.

[24] C. Lamini, S. Benhlima, and A. Elbekri, "Genetic algorithm based approach for autonomous mobile robot path planning," *Procedia Computer Science*, vol. 127, pp. 180–189, 2018.

[25] R. M. C. Santiago, A. L. De Ocampo, A. T. Ubando, A. A. Bandala, and E. P. Dadios, "Path planning for mobile robots using genetic algorithm and probabilistic roadmap," in *2017IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM)*. IEEE, 2017, pp. 1–5.

[26] X. Liang, P. Jiang, and H. Zhu, "Path planning for unmanned surface vehicle with dubins curve based on ga," in *2020 Chinese Automation Congress (CAC)*. IEEE, 2020, pp. 5149–5154.

[27] C. Ma, R. He, and W. Zhang, "Path optimization of taxi carpooling," *PLoS One*, vol. 13, no. 8, p. e0203221, 2018.

[28] S. Luo, Y. Singh, H. Yang, J. H. Bae, J. E. Dietz, X. Diao, and B.-C. Min, "Image processing and model-based spill coverage path planning for unmanned surface vehicles," in *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, 2019, pp. 1–9.

[29] F. M. Santa, S. O. Rivera, and M. A. Saavedra, "Enfoque de navegación global para un robot asistente," *Tecnura*, vol. 21, no. 51, p. 105, 2017.

[30] J. Krejsa and S. Vechet, "Determination of optimal local path for mobile robot," in *Mechatronics 2017*, T. Březina and R. Jabłoński, Eds. Cham: Springer International Publishing, 2018, pp. 637–643.

[31] A. Barrero, M. Robayo, and E. Jacinto, "Algoritmo de navegación a bordo en ambientes controlados a partir de procesamiento de imágenes," *Tekhnê*, vol. 12, no. 2, pp. 23–34, 2015.

[32] F. Martinez *et al.*, "Navigable points estimation for mobile robots using binary image skeletonization," in *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225. International Society for Optics and Photonics, 2017, p. 1022505.

[33] N. B. A. Latip, R. Omar, and S. K. Debnath, "Optimal path planning using equilateral spaces oriented visibility graph method," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 6, p. 3046, 2017.

[34] J. D. Contreras, F. Martínez *et al.*, "Path planning for mobile robots based on visibility graphs and a* algorithm," in *Seventh International Conference on Digital Image Processing (ICDIP 2015)*, vol. 9631. SPIE, 2015, pp. 345–350.

[35] S. B. Mane and S. Vhanale, "Genetic algorithm approach for obstacle avoidance and path optimization of mobile robot," in *Computing, Communication and Signal Processing*, B. Iyer, S. Nalbalwar, and N. P. Pathak, Eds. Singapore: Springer Singapore, 2019, pp. 649–659.

Cryptanalysis of a Hamming Code and Logistic-Map based Pixel-Level Active Forgery Detection Scheme

Oussama Benrhouma
Faculty of Computer and Information Systems
Islamic University of Medinah, Medinah, KSA

***Abstract*—In this paper, we analyze the security of a fragile watermarking scheme for tamper detection in images recently proposed by S. Prasad et al. The chaotic functions are used in the scheme to exploit its pseudo-random behavior and its sensibility to initial condition and control parameter, but despite that, security flaws have been spotted and cryptanalysis of the scheme is conducted. Experimental results shows that the scheme could not withstand the attack and watermarked images were manipulated without triggering any alarm in the extraction scheme. In this paper, two different approaches of attacks are demonstrated and conducted to break the scheme. This work falls into the context of improving the quality of the designed cryptographic schemes taking into account several cryptanalysis techniques.**

***Keywords*—Cryptanalysis; watermarking; tamper detection; attack; chaotic functions; forgery localization**

I. Introduction

Nowadays we are living in the era of technology, and with a huge leap of internet technology the advancement is going faster and faster thanks to the easy and fast exchange of information, this makes led to the emergence of powerful software and hardware. Powerful devices with huge computational capacity became available at reasonable prices.

The amount of data exchanged via the internet is huge, multimedia contents represent a big percentage of these files, and with the presence of powerful and easy to use software the manipulation of these files became easier. With more than 300 million images uploaded every single day, the protection of these images became a necessity since it can be used to spread fake news, create problems between individuals or even nations, and now digital images could be presented as evidence in courtrooms. for these reasons, the scientific community is facing the challenge to present efficient solutions to control the integrity of these images.

Digital watermarking present a solution for these problems [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Digital watermarking could be classified into three categories: robust, fragile, and semi-fragile watermarking schemes.

Robust watermarking schemes are typically designed for copyright protection [20], [21], [22]. The owner should be able to extract and verify an embedded watermark even from a falsified image, on the other hand, Fragile and semi-fragile watermarking schemes are designed to control the integrity of the cover image [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [18], [19], any unauthorized modification on the watermarked image should affect the embedded watermark and therefore detected by the legitimate receiver. The legitimate receiver is typically whoever possesses the secret key(s) to extract the watermark, and despite that the schemes are protected by secret key(s), successful attacks on these schemes has been conducted and the watermarks has been removed without possession of the key(s) [33], [34], [35], [36], [37]. The work presented by the cryptanalysts helped improving the quality of the future proposed security schemes. In this context w analyze the security of a recently proposed fragile watermarking scheme by S. Prasad et al [1], security flaws have been spotted in the scheme and two different types of attacks are performed and we were able to modify the image without being detected by the detection scheme, finally, an improvement of the scheme is proposed to cover the security problems.

The rest of the paper is organized as follows: In Section 2 we present a description of the scheme under study, Section 3 two types of attack are demonstrated and results are presented, finally, the paper is concluded in Section 4.

II. The Scheme under Study

The scheme in [1] proposes a fragile watermarking scheme for tamper detection in digital images. The scheme is based on (7,4) hamming code and logistic map: for each pixel the 4 most significant bits (MSBs) are selected and (7,4) hamming code is used to generate 3-bits authentication code that is then further processed using the logistic map and embedded into the LSBs of the pixel in question. In this section we present a brief description of the scheme under study.

A. Authentication Watermark Generation and Embedding

Given a cover image I with size $(M \times N)$ the steps leading to the generation and the embedding of the watermark are as follows:

Step 1: The logistic map is used to generate a pseudo-random sequence α where $\alpha = \{\alpha_i; i = 1 : (M \times N)\}$.
The Logistic map is defined by equation 1.The values generated by the equations are in [0,1], α_0 represents the initial condition provided by the user, and β is the control parameter of the function, where $\beta \in [0, 4]$.

$$\alpha_{i+1} = \beta\alpha_i(1 - \alpha_i) \quad (1)$$

The initial condition α_0 and the control parameter β are considered as secret keys of scheme.

Step 2: At this point we have a pseudo-random sequence $\alpha = \alpha_i \quad (i = 1 : MN)$, with the same size of

the image, each value from the sequence α will be associated to a pixel, where i represent the index of the pixel in processing.
The pseudo-random sequence is then converted to be in the range from 0 to 7 using the equations 2, 3 and 4.

$$A_i = \alpha_i \times 255 \tag{2}$$

$$B_i = round(A_i) \tag{3}$$

$$K_i = mod(B_i, 8) \tag{4}$$

Step 3: The i^{th} pixel in the cover image I is selected, converted to binary then its 4 MSBs are selected to compute its hamming code $c = (c_7, c_6, c_5, c_4, c_3, c_2, c_1)$.
The watermark is considered the 3 LSBs of the calculated hamming code: $W = (c_3, c_2, c_1)$

Step 4: The computed watermark is converted into an integer to obtain T.

Step 5: Starting from the secret value K_i a list R is created:

$$R = \{K_i, (K_i + 1)\ mod\ 8, (K_i + 2)\ mod\ 8, ..., (K_i + 7)\ mod\ 8, \} \tag{5}$$

Step 6: The value of the watermark T is Searched within the list R and its position in R is saved as "j".

Step 7: Calculate $z = mod\,(P_i, 8)$. Where P_i is the pixel in processing.

Step 8: Calculate list PR, where $PR = \{P_i - z + t\,;\, t = 0, 1, 2..., 7\}$.

Step 9: The watermarked pixel is represented by the j^{th} element in the list PR.

Step 10: The rest of the cover image is processed by applying the steps 3 to 9 to obtain the watermarked image WI.

A flowchart of the embedding schemes is shown in Fig. 1.

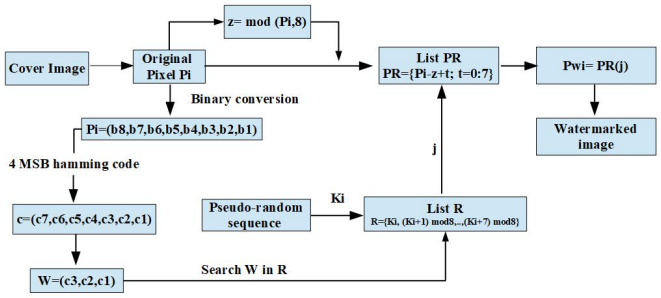


Fig. 1. Flowchart of the Embedding Phase [1]

B. Extraction and Tamper Detection

Given a received watermarked image WI. The steps leading to the extraction of the watermark in order to locate any possible tampering in the image are described as follows:

Step 1: Generate the same pseudo-random sequence α using the logistic map defined in equation 1 with the parameters α_0 and β as secret keys keys.
$\alpha = \{\alpha_i\ ;\ i = 1 : (M \times N)\}$. where $(M \times N)$ is the size of the image WI.

Step 2: The pseudo-random sequence α is then converted to be in the range from 0 to 7 using the equations 2, 3 and 4.
The list K with the same size as the image and each element represents the secret value that will be used to generate the list R for each pixel.

Step 3: The i^{th} pixel P_{Wi} in the received image WI is selected, then converted to binary then its 4 MSBs are selected to compute its hamming code $c = (c_7, c_6, c_5, c_4, c_3, c_2, c_1)$.
The 3-bits authentication code watermark is the 3 LSBs of the calculated hamming code c: $W = (c_3, c_2, c_1)$

Step 4: The list R is generated starting from the elements of the list K: for the i^{th} pixel the element K_i is used to calculate the list R:

$$R = \{K_i, (K_i + 1)\ mod\ 8, (K_i + 2)\ mod\ 8, ..., (K_i + 7)\ mod\ 8, \} \tag{6}$$

Step 5: Compute $z = mod\,(P_{Wi}, 8) + 1$ which represents the index of the extracted watermark E_{AC} in the list R.

Step 6: The comparison between the extracted watermark E_{AC} and the calculated one W will reveal if the pixel in question has been tampered with: each pixel where $E_{AC} \neq W$ is considered falsified, therefore its position in the received image is set to zero which represent the black color.

A flowchart of the extraction and tamper detection schemes is shown in Fig. 2.

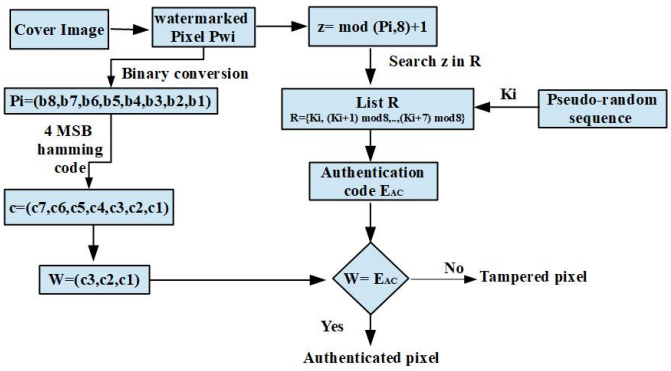


Fig. 2. Flowchart of the Extraction Phase [1]

III. Cryptanalysis of the Scheme

A. Offline Attack

In general, an attacker's goal is is either to guess or recover the value of the secret key(s) or something equivalent to the key(s) in order to recover the plaintext without knowledge of the secret key and that is due to kerckhoff's principle that states

that everything about the cryptosystem is public knowledge except for the keys.

In other words, the only thing secret about a cryptosystem is the secret key(s), everything else should be known and the job of a cryptographer is to design a cryptosystem that stands against any type of attack taking into consideration Kerckhoff's law [38].

The scheme under study [1] is a fragile watermarking system for tamper detection in digital images, after a successful cryptanalysis we should be able to manipulate the watermarked images without being detected by the extraction scheme. To achieve that goal, the keys or the equivalent of the keys are needed.

In the scheme in [1] the keys are the initial condition α_0 and the control parameter β of the logistic map.

The keys $(\alpha_0, beta)$ are used to generate a pseudo-random sequence α with the same size of the image then the sequence is quantified to be in the range of [0,7] to obtain the sequence K ad each element K_i in K is assigned to the pixel i in the image and the sequence R is constructed : $R = \{K_i, (K_i + 1)mod\,8, (K_i + 2)mod\,8, ..., (K_i + 7)mod\,8\}$

One of the main features of the chaotic maps is the high sensibility to initial conditions and control parameter, which make the attempt of any prediction or guess to their values starting from the pattern of the function nearly impossible, beside the pattern of the function is not available, but we know that it has been used to construct the lists K and R.

Since that the main keys are very hard to find our goal is to reveal alternative keys which are the lists R for each pixel we attempt to modify, and the list K if needed in any other attack intercepted from the same source.

In this section we will demonstrate how to reveal the list R for each pixel and as a result we will be able to construct the list K for the image:

Given an intercepted watermarked image "WI" with size $M \times N$, the steps leading to the revelation of the lists R and K are as follows:

Step 1: The i^{th} pixel P_{Wi} in the intercepted watermarked image WI is selected, then converted to binary then its 4 MSBs are selected to compute its hamming code $c = (c_7, c_6, c_5, c_4, c_3, c_2, c_1)$.
The 3-bits authentication code watermark is the 3 LSBs of the calculated hamming code c: $W = (c_3, c_2, c_1)$
It should be noted that i represents the index of the pixel in the image : $i = 1 : M * N$.

Step 2: Compute $z = mod\,(P_{Wi}, 8)$ which represents the index of the watermark W in the list R.

Step 3: Starting from the z^{th} position, the list R could be reconstructed using equation 7.

$$R((z+j)mod8) = (W+j)mod8 \quad where j = 0 : 7. \tag{7}$$

It should be noted that the authors in [1] used $z = mod\,(P_{Wi}, 8) + 1$ based on the indexation starts from 1 not 0, in our attack we dealt with the lists from 0 to 7 indexation.

Step 4: The first element in the list R represents the value K_i in the pseudo random-sequence K.
Once all pixels of the intercepted image are processed the pseudo-random sequence K is revealed.

Step 5: The i^{th} pixel could now be modified and the watermark is substituted with the new one with the possession of the list R.

With the possession of the equivalent keys (The lists R and K), the watermarked image could now be manipulated and the watermark is replaced without being detected by the extraction scheme. Next we present two examples how to calculate the list R for a given pixel and find the corresponding value K_i.

a) Example 1:: In the first example the value of the pixel $P_i = 165$ and $K_i = 3$.

First the watermark embedding process:

1) The list R is constructed using equation 5

$$R = \{K_i, (K_i+1)mod8, (K_i+2)mod8, ..., (K_i+7)mod8, \}$$

$$\Rightarrow R = \{3, 4, 5, 6, 7, 0, 1, 2\}$$

2) Calculate $z = mod\,(P_i, 8) \quad \Rightarrow z = mod(165, 8) = 5$.

3) Calculate list PR, where $PR = \{P_i - z + t\,;\, t = 0, 1, 2..., 7\}$.
$\Rightarrow PR = \{160, 161, 162, 163, 164, 165, 166, 167\}$.

4) P_i is converted to binary and the hamming code c for its 4 MSBs is calculated:
$\Rightarrow (165)_{10} = (10100101)_2$
$\Rightarrow c = H_{(4,7)}(1010) = 1010010 \Rightarrow T = (c_3, c_2, c_1) = (010)_2 = 2_{10}$.

5) Search for T_{10} in R and and its position in PR is considered as watermarked pixel: $R = \{3, 4, 5, 6, 7, 0, 1, 2\}; \quad T = 2_{10} \Rightarrow 7^{th} position$
$PR = \{160, 161, 162, 163, 164, 165, 166, 167\} \Rightarrow P_{Wi} = PR(7) = 167$

The image is then intercepted during transmission, next we demonstrate how the list R is calculated along with the value K_i.

1) $P_{Wi} = 167$ is the value of the ith pixel in the intercepted watermarked image "WI"
P_{Wi} is then converted to binary and the hamming code is calculated for its 4 MSBs to obtain T.
$\Rightarrow (167)_{10} = (10100111)_2$
$\Rightarrow c = H_{(4,7)}(1010) = 1010010 \Rightarrow T = (c_3, c_2, c_1) = (010)_2 = 2_{10}$.

2) Calculate $z = mod\,(P_{Wi}, 8) \Rightarrow z = mod(167, 8) = 7$.
This means that the z^{th} position in R contain the value of T.
$\Rightarrow R(7) = 2; \Rightarrow R = \{?, ?, ?, ?, ?, ?, ?, 2\}$

3) The list R is now constructed using equation 7.
$R((z+j)\,mod8) = (W+j)\,mod8 \quad where j = 0 : 7$
$\Rightarrow R(7) = 2; R(0) = 3; R(1) = 4; R(2) = 5; R(3) = 6; R(4) = 7; R(5) = 0; R(6) = 1$
$\Rightarrow R = \{3, 4, 5, 6, 7, 0, 1, 2\}$

4) The first element in the list R represents the value K_i where i is the index of the pixel in question : $K_i = 3$.
5) The i^{th} pixel could now be modified and the watermark substituted with the possession of the list R.

b) Example 2:: In the second example the value of the pixel $P_i = 99$ and $K_i = 5$.

First the watermark embedding process:

1) The list R is constructed using equation 5

$$R = \{K_i, (K_i+1)mod8, (K_i+2)mod8, ..., (K_i+7)mod8, \}$$

$\Rightarrow R = \{5, 6, 7, 0, 1, 2, 3, 4\}$
2) Calculate $z = mod(P_i, 8) \quad \Rightarrow z = mod(99, 8) = 3$.
3) Calculate list PR, where $PR = \{P_i - z + t\ ;\ t = 0, 1, 2..., 7\}$.
$\Rightarrow PR = \{96, 97, 98, 99, 100, 101, 102, 103\}$.
4) P_i is converted to binary and the hamming code c for its 4 MSBs is calculated:
$\Rightarrow (99)_{10} = (01100011)_2$
$\Rightarrow c = H_{(4,7)}(0110) = 0110011 \Rightarrow T = (c_3, c_2, c_1) = (011)_2 = 3_{10}$.
5) Search for T_{10} in R and and its position in PR is considered as watermarked pixel:
$R = \{5, 6, 7, 0, 1, 2, 3, 4\}; \quad T = 2_{10} \Rightarrow 6^{th} position$
$PR = \{96, 97, 98, 99, 100, 101, 102, 103\} \Rightarrow P_{Wi} = PR(6) = 102$

The image is then intercepted during transmission, next we demonstrate how the list R is calculated along with the value K_i.

1) $P_{Wi} = 102$ is the value of the i^{th} pixel in the intercepted watermarked image "WI"
P_{Wi} is then converted to binary and the hamming code is calculated for its 4 MSBs to obtain T.
$\Rightarrow (102)_{10} = (01100110)_2$
$\Rightarrow c = H_{(4,7)}(0110) = 0110011 \Rightarrow T = (c_3, c_2, c_1) = (011)_2 = 3_{10}$.
2) Calculate $z = mod\ (P_{Wi}, 8) \Rightarrow z = mod(102, 8) = 6$.
This means that the z^{th} position in R contain the value of T.
$\Rightarrow R(6) = 3; \Rightarrow R = \{?, ?, ?, ?, ?, ?, 3, ?\}$
3) The list R is now constructed using equation 7.
$R((z+j)\,mod 8) = (W+j)\,mod 8 \quad where\, j = 0:7$
$\Rightarrow R(6) = 3; R(7) = 4, R(0) = 5; R(1) = 6; R(2) = 7; R(3) = 0; R(4) = 1; R(5) = 2$
$\Rightarrow R = \{5, 6, 7, 0, 1, 2, 3, 4\}$
4) The first element in the list R represents the value K_i where i is the index of the pixel in question : $K_i = 5$.
5) The i^{th} pixel could now be modified and the watermark substituted with the possession of the list R.

B. Online Attack

The second approach to attack the scheme under study is to use one of the online attacks. Online attacks could be summarized in three main approaches [39]:

1) **KPA** Known plaintext attack : In this scenario the cryptanalyst has one or several plain-text and their corresponding cipher-text. the cryptanalyst then tries to conclude the key or an equivalent key from the analysis of these pairs.
2) **CPA** Chosen plain-text attack: As in the case of KPA the cryptanalyst possesses pairs of plain-text and their corresponding ciphers only in this scenario, the attacker has access to the encryption machinery and can chose the plain-texts to be encrypted.
3) **CCA** Chosen cipher-text attack : In this scenario the attacker has access to the decryption machinery and can chose cipher-texts to get the corresponding plain-texts. Based on the study of these plain/cipher-texts the cryptanalyst tries to conclude the key or an equivalent of the key.

These scenarios represent the most common techniques in cryptanalysis. any security system should be tested to avoid vulnerability against these attacks.

Using KPA or CPA, only a single pair of original image and its corresponding watermarked image is needed to break the system and reveal the secret keys (The list R for each pixel and the list K for the image):

Let "OI" be the original image and "WI" its corresponding watermarked image with size $M \times N$, and OP_i, WP_i are the pixels of "OI" and "WI" respectively, where i represents the index of the pixel, the secret lists R and K could be calculated as follows:

1) Find $z = mod\,(OP_i, 8)$, then calculate the list PR, where $PR = \{P_i - z + t\ ;\ t = 0, 1, 2..., 7\}$.
2) Convert OP_i (or WP_i) to binary and and calculate the hamming code for its 4 MSBs : $c = (c_7, c_6, c_5, c_4, c_3, c_2, c_1)$. The value T is the integer value of $W = (c_3, c_2, c_1)$
3) Find the value of WP_i in the list PR and save its index as j.
4) j represents the position of T in the list R, so the list R could now be calculated using equation 8

$$R((j+t)\,mod 8) = (T+t)\,mod 8 \quad where\, t = 0:7. \tag{8}$$

5) The first element in the list R represents the secret value K_i. Once all pixel are processed the list K will be revealed.

With the revelation of the secret list K the image could be manipulated and the watermark is successfully replaced without being detected by the extraction scheme.

a numerical example is presented next :

a) Example:: In this example the original value of the pixel $OP_i = 165$ and its corresponding watermarked pixel $WP_i = 167$.

1) Calculate $z = mod\ (OP_i, 8) \Rightarrow z = mod(165, 8) = 5$.
Calculate list PR, where $PR = \{P_i - z + t\ ;\ t = 0, 1, 2..., 7\}$.
$\Rightarrow PR = \{160, 161, 162, 163, 164, 165, 166, 167\}$.
2) OP_i is converted to binary and the hamming code c for its 4 MSBs is calculated:

$\Rightarrow (165)_{10} = (10100101)_2$
$\Rightarrow c = H_{(4,7)}(1010) = 1010010 \Rightarrow T = (c_3, c_2, c_1) = (010)_2 = 2_{10}$.

3) Search for WP_i in the list PR. $\Rightarrow$ 167 in the 7^{th} position. $\Rightarrow j = 7$.
4) $R(j) = T \Rightarrow R(7) = 2$,$\Rightarrow R(7) = 2; \Rightarrow R = \{?, ?, ?, ?, ?, ?, ?, 2\}$
Starting from the position $j = 7$ the list R is revealed using equation 8.
$R((j+t) \, mod \, 8) = (T+t) \, mod \, 8 \quad where \, t = 0 : 7.$
$\Rightarrow R(7) = 2; R(0) = 3; R(1) = 4; R(2) = 5; R(3) = 6; R(4) = 7; R(5) = 0; R(6) = 1$
$\Rightarrow R = \{3, 4, 5, 6, 7, 0, 1, 2\}$
5) The first element in R represents the i^{th} element in the secret list K: $K_i = R(0) = 3$.

Fig. 3 shows the results of the attack. Multiples images were used in the experiments, we were able to calculate the keys used in the embedding process, as a result, the watermarks were successfully removed in order to manipulate the images, then using the calculated keys, new watermark are embedded into the falsified images in order to prevent any alarms in the extraction process.

The experiments shows that the extraction scheme failed to detect the falsifications which proves the weakness of the proposed scheme.

IV. Conclusion

In this paper, a cryptanalysis of a recently proposed watermarking scheme is conducted, two types of attacks were conducted successfully. As a result, the watermarked images could be falsified without triggering any alarm in the extraction process. This proves that even if very complicated steps were used in the design of a cryptographic scheme, that doesn't mean that the scheme is secure, several cryptalysis techniques could be used to attack these scheme, and these cryptanalysis techniques should be taken into consideration when designing a cryptographic scheme. As future work, an improvement of the attacked scheme could be proposed to cover the flaws and problems demonstrated in this paper.

Acknowledgment

The authors would like to thank the deanship of research at the Islamic University of Madinah, Kingdom of Saudi Arabia for supporting this research.

References

[1] S. Prasad and A. K. Pal, "Hamming code and logistic-map based pixel-level active forgery detection scheme using fragile watermarking," *Multimedia Tools and Applications*, Apr 2020. [Online]. Available: https://doi.org/10.1007/s11042-020-08715-x

[2] N. Sivasubramanian and G. Konganathan, "A novel semi fragile watermarking technique for tamper detection and recovery using iwt and dct," *Computing*, 2020.

[3] O. Evsutin and K. Dzhanashia, "Watermarking schemes for digital images: Robustness overview," *Signal Processing: Image Communication*, vol. 100, p. 116523, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596521002551

[4] W. H. Alshoura, Z. Zainol, J. S. Teh, and M. Alawida, "An fpp-resistant svd-based image watermarking scheme based on chaotic control," *Alexandria Engineering Journal*, vol. 61, no. 7, pp. 5713–5734, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016821007213

[5] M. Swain and D. Swain, "An effective watermarking technique using btc and svd for image authentication and quality recovery," *Integration*, vol. 83, pp. 12–23, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167926021001255

[6] S. O. V. Aslantas and S. Ozturk, "Improving the performance of dct-based fragile watermarking using intelligent optimization algorithms," *Opt Commun*, vol. 282, pp. 2806–2817, 2009.

[7] G. Bhatnagar and B. Raman, "A new robust reference logo watermarking scheme," *Multimedia Tools Appl*, vol. 52, pp. 621–640, 2011.

[8] Y. Y. P.P. Niu, X.Y. Wang and M. Lu, "A novel color image watermarking scheme in nonsampled contourlet-domain," *Expert Syst Appl*, vol. 38, pp. 2081–2098, 2011.

[9] C.-Y. Lin and S.-F. Chang, "Sari: Self-authentication-and-recovery image watermarking system," *ACM Multimedia*, vol. Ottawa, Canada, Sep. 30 - Oct 5, 2001.

[10] T.-Y. Lee. and S. D. Lin, "Dual watermark for image tamper detection and recovery," *Pattern Recognition*, vol. 41, pp. 3497–3506, 2008.

[11] S. Poonkuntran and R. S. Rajesh, "Chaotic model based semi fragile watermarking using integer transforms for digital fundus image authentication," *Multimed Tools Appl*, vol. DOI 10.1007/s11042-012-1227-5, 2012.

[12] H. H. Yaoran Huo and F. Chen, "A semi-fragile image watermarking algorithm with two-stage detection," *Multimed Tools Appl*, vol. DOI 10.1007/s11042-012-1317-4, 2013.

[13] T. Luo, G. Jiang, X. Wang, M. Yu, F. Shao, and Z. Peng, "Stereo image watermarking scheme for authentication with self-recovery capability using inter-view reference sharing," *Multimed Tools Appl*, vol. DOI 10.1007/s11042-013-1435-7, 2013.

[14] S.-J. Horng, M. Farfoura, P. Fan, X. Wang, T. Li, and J.-M. Guo, "A low cost fragile watermarking scheme in h.264/avc compressed domain," *Multimedia Tools and Applications*, vol. 72, pp. 2469–2495, 2014.

[15] S.-J. Horng, D. Rosiyadi, P. Fan, X. Wang, and M. Khan, "An adaptive watermarking scheme for e-government document images," *Multimedia Tools and Applications*, vol. 72, pp. 3085–3103, 2014.

[16] S.-J. Horng, D. Rosiyadi, T. Li, T. Takao, M. Guo, and M. K. Khan, "A blind image copyright protection scheme for e-government," *Journal of Visual Communication and Image Representation*, vol. 24, pp. 1099 – 1105, 2013.

[17] D. Rosiyadi, S.-J. Horng, P. Fan, X. Wang, M. Khan, and Y. Pan, "Copyright protection for e-government document images," *MultiMedia, IEEE*, vol. 19, pp. 62–73, 2012.

[18] O. Benrhouma, H. Hermassi, and S. Belghith, "Tamper detection and self-recovery scheme by dwt watermarking," *Nonlinear Dynamics*, vol. 79, no. 3, pp. 1817–1833, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11071-014-1777-3

[19] O. Benrhouma, H. Hermassi, A. A. Abd El-Latif, and S. Belghith, "Chaotic watermark for blind forgery detection in images," *Multimedia Tools and Applications*, pp. 1–24, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11042-015-2786-z

[20] W. Lu, H. Lu, and F. Chung, "Robust digital image watermarking based on sub-sampling." *Applied Mathematics and computation*, vol. 181, pp. 886–893, 2006.

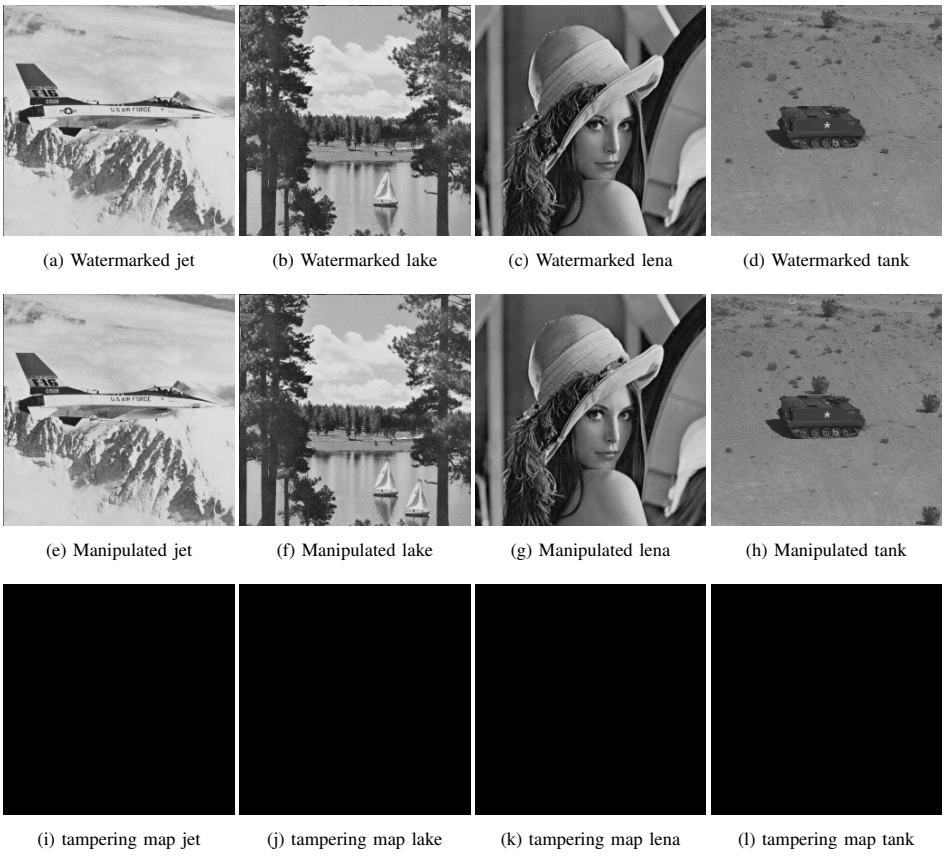
[21] D. Simitopoulos, D. Koutsonanos, and M. Strintzis, "Robust image watermarking based on generalized radon transformations." *IEEE Transctions on Circuits and Systems for Video Technology.*, vol. 13, pp. 732–745, 2003.

[22] H. Tang, C.W. Hang, "A feature-based robust digital image watermaking shceme." *IEEE Transctions on Signal Processing.*, vol. 51, pp. 950–958, 2003.

[23] M. Celik, G. Sharmar, and A. Tekalp, "Hierarchical watermarking for secure image authentication with localization." *IEEE Transctions on Image Processing.*, vol. 11, pp. 585–594, 2002.

[24] E. chang, M. Kankanhalli, X. Guan, Z. Huang, and Y. Wu, "Robust image authentication using content based compression." *ACM Multimedia System Journal*, vol. 2, pp. 121–130, 2003.

[25] J. Fridrich, M. Goljan, and A. Baldoza, "New fragile authentication watermarks for image." *Proceeding of IEEE International Conference on Image Processing.*, vol. 1, pp. 446–449, 2000.



(a) Watermarked jet (b) Watermarked lake (c) Watermarked lena (d) Watermarked tank

(e) Manipulated jet (f) Manipulated lake (g) Manipulated lena (h) Manipulated tank

(i) tampering map jet (j) tampering map lake (k) tampering map lena (l) tampering map tank

Fig. 3. Results of the Attack: Tampered Images and the Corresponding Tampering Map

[26] C. Lin and S. Chang, "A robust image authentication method distinguishing jpeg compression from malicious manipulation." *IEEE Transctions on Circuits and Systems for Video Technology.*, vol. 11, pp. 153–168, 2001.

[27] A. Paquet, R. Ward, and I. Pitas, "Wavelet packets-based digital watermaking for image verification and authentication." *Signal Processing.*, vol. 183, pp. 2117–2132, 2003.

[28] N. Wong, P. Memon, "Secret and public key authentication watermarking schemes that resist vector quantization attack." *Proceeding ofSPIE on Security and Watermarking of Multimedia Contents.*, vol. 3971, pp. 417–427, 2000.

[29] F. Yeung, M. Mintzer, "An invisible watermarking technique for image verification." *Proceeding of IEEE International Conference on Image Processing.*, vol. 2, pp. 680–683, 1997.

[30] J. Fridrich, "Security of fragile authentication watermaks with localization." *Proceeding ofSPIE on Security and Watermarking of Multimedia Contents.*, vol. 4675, pp. 691–700, 2002.

[31] B. B. Haghighi, A. H. Taherinia, and A. H. Mohajerzadeh, "Trlg: Fragile blind quad watermarking for image tamper detection and recovery by providing compact digests with optimized quality using lwt and ga," *Information Sciences*, vol. 486, pp. 204 – 230, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025519301707

[32] J. Molina-Garcia, B. P. Garcia-Salgado, V. Ponomaryov, R. Reyes-Reyes, S. Sadovnychiy, and C. Cruz-Ramos, "An effective fragile watermarking scheme for color image tampering detection and self-recovery," *Signal Processing: Image Communication*, vol. 81, p. 115725, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596519306897

[33] L. Teng, X. Wang, and X. Wang, "Cryptanalysis and improvement of a chaotic system based fragile watermarking scheme," {*AEU*} - *International Journal of Electronics and Communications*, vol. 67, no. 6, pp. 540 – 547, 2013.

[34] M. Botta, D. Cavagnino, and V. Pomponiu, "A successful attack and revision of a chaotic system based fragile watermarking scheme for image tamper detection," {*AEU*} - *International Journal of Electronics and Communications*, vol. 69, no. 1, pp. 242 – 245, 2015.

[35] M. Li, J. Zhang, and W. Wen, "Cryptanalysis and improvement of a binary watermark-based copyright protection scheme for remote sensing images," *Optik - International Journal for Light and Electron Optics*, vol. 125, no. 24, pp. 7231 – 7234, 2014.

[36] O. Benrhouma, H. Hermassi, and S. Belghith, "Security analysis and improvement of an active watermarking system for image tampering detection using a self-recovery scheme," *Multimedia Tools and Applications*, pp. 1–24, 2016.

[37] H. He and J. Zhang, "Cryptanalysis on majority-voting based self-recovery watermarking scheme," *Telecommun Syst*, vol. 49, pp. 231–238, 2012.

[38] A. Kerckhoffs, "La cryptographie militaire," *Journal des sciences militaires*, vol. 9, pp. 5–38, 1883.

[39] D. A. Guardeno, "Framework for the analysis and design of encryption strategies based on discrete-time chaotic dynamical systems," Ph.D. dissertation, Escuela Técnica Superior de Ingenieros Agronomos, Universidad Politécnica de Madrid, 2009.

Wifi Indoor Positioning with Genetic and Machine Learning Autonomous War-Driving Scheme

Pham Doan Tinh
School of Electrical and Electronics Engineering
Hanoi University of Science and Technology
No.1 Dai Co Viet street, Hanoi, Vietnam

Bui Huy Hoang
School of Electrical and Electronics Engineering
Hanoi University of Science and Technology
No.1 Dai Co Viet street, Hanoi, Vietnam

***Abstract*—Wifi Fingerprinting is a widely used method for indoor positioning due to its proven accuracy. However, the offline phase of the method requires collecting a large quantity of data which costs a lot of time and effort. Furthermore, interior changes in the environment can have impact on system accuracy. This paper addresses the issue by proposing a new data collecting procedure in the offline phase that only needs to collect some data points (Wi-fi reference point). To have a sufficient amount of data for the offline phase, we proposed a genetic algorithm and machine learning model to generate labeled data from unlabeled user data. The experiment was carried out using real Wi-fi data collected from our testing site and the simulated motion data. Results have shown that using the proposed method and only 8 Wi-fi reference points, labeled data can be generated from user's live data with a positioning error of 1.23 meters in the worst case when motion error is 30%. In the online phase, we achieved a positioning error of 1.89 meters when using the Support Vector Machine model at 30% motion error.**

***Keywords*—Wifi fingerprinting; indoor positioning; machine learning; genetic algorithm**

I. Introduction

As people spend more time indoors, many location-based applications and services require to known user indoor location. While the global positioning system (GPS) is a popular positioning method, it can hardly be applied to indoor environments because lacking in line of sight (LOS). Therefore, many indoor positioning techniques have emerged and been proposed in recent years.

An approach that many researchers have taken is to use the network infrastructure like Wi-fi [1], [2], Bluetooth [3], [4], Zigbee [5], Ultra-Wideband [6] to perform indoor localization. Wi-fi is the most common because it is likely to be installed in most public indoor places like malls, stations, airports and nearly every smartphone is equipped with a Wi-fi transceiver module. A popular method that utilizes Wi-fi signal is Wifi Receive Signal Strength Indicator (RSSI) Fingerprinting [7], [8], [9], [10]. The approach assumed that the RSSI measurement from access points (APs) for every location is unique. For that reason, RSSI measurements are recorded and stored in a database called the radio map. Whenever a new RSSI measurement is generated by the user, it will be matched to the similar one in the database. One of the popular matching algorithms is The Nearest Neighbors in Signal Space (NNSS) [11], which calculates the distance of signal space between the observed data and the recorded data. The step of constructing the database is also known as the offline phase and the matching step is known as the online phase. In the matching step, machine learning can be also be used to take advantage of the powerful pattern recognition ability to produce more accurate results [10].

Another popular approach focus on exploiting the inertial measuring units (IMU) because of their availability in mobile devices and fast measurement update. A well-known method that takes advantage of the IMU measurements is Pedestrian Dead Reckoning (PDR) [12], [13], [14]. PDR extract step event [15], step length [16] and heading angle [17] from IMU raw measurements and output the location of the user. The major benefit of this approach is that no infrastructure needs to be installed. In addition, measurements are regularly updated which enables real-time localization. However, the initial position of the target is required to be predetermined in PDR because PDR can not locate the current position of the target without the knowledge of the target's previous location in the environment. Furthermore, IMU sensors are subjected to noise, interference and disturbance thus produces accumulated errors over time. For that reason, sensor fusion methods like Kalman Filter [18] or Madgwick Filter [19] were employed to counter the error but due to the complexity of those filters, they raise the amount of computation to solve the indoor localization problem.

Wi-fi Fingerprinting has advantages when it comes to precision and deployment cost compared with other indoor positioning schemes. However, it is worth mentioning the drawbacks of the method. First, Wi-fi Fingerprinting method performance is suffered from multipath, shadowing, and interference in the environment [20]. Secondly, the number of data points in the radio map can affect the positioning accuracy. Therefore, in the offline phase of conventional Wi-fi Fingerprinting, the coordinate system of the whole area needs to be built, then numerous data points are marked to have their RSSI measurements collected. This procedure consumes a lot of time and effort. Additionally, changes like adding or moving objects in the environment could make the radio map no longer reliable and have to be reconstructed to maintain accuracy. While the offline phase of the Fingerprinting method poses difficulties in collecting data, the PDR method creates a lot of data but they are unlabeled. Aiming to reduce the amount of data needed to be collected in the offline phase of the Wi-fi Fingerprinting method and taking advantage of the results from the PDR method, this paper proposed a new architecture for the offline phase of the Wi-fi Fingerprinting method. From the results of PDR, a genetic algorithm is

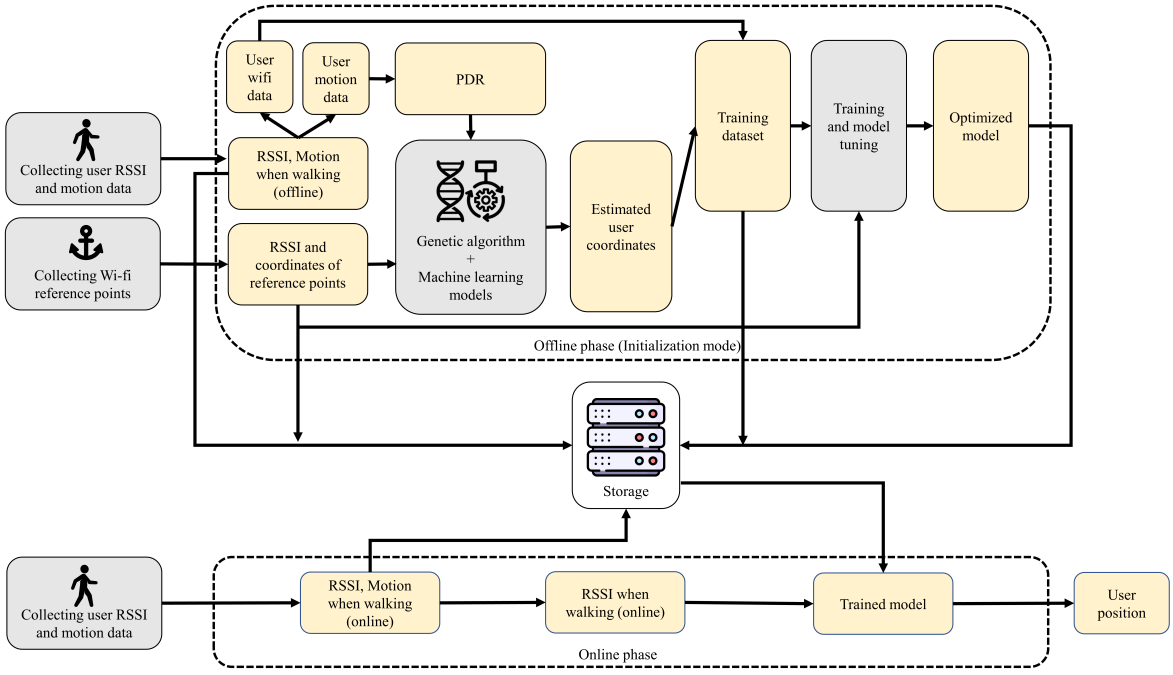


Fig. 1. Offline Phase Initialization Mode and Online Phase of the Proposed Method

implemented to combined with machine learning algorithms to find labels (locations) for the user unlabeled Wifi RSSI data. By creating more labeled data from user data, the proposed method reduces the spent time and effort to collect labeled data while still achieve good results.

Several approaches that used motion sensors and Wi-fi signals for indoor positioning have been studied in the past. Wi-fi SLAM proposed by Brian Ferris et al. in [21] take advantage of a technique called Simultaneous Localization and Mapping (SLAM), which is a popular navigation method in an unknown environment. The authors used Gaussian Process Latent Variable Model to find the location of unlabeled signal strength data in the latent space and combine it with motion data to rebuild the topological connectivity graph to perform indoor localization. Naguib in [22] proposed an indoor positioning scheme that combines multi sources of information, which are motion sensors and Wi-fi using a low-complexity version of particle filter to increase accuracy. Constandache in [23] utilized electronic compass and accelerometers in phones to measure user speed and orientation. Then the recorded data is matched against possible path signatures in the local electronic map. The proposed method is different from others because the user motion data are not used together with Wi-fi to directly predict the user location but they are only utilized in the offline phase to generate more labeled data for the machine learning model. As the motion data is only utilized in the offline phase, the proposed method does not contain any heavy computation in the online phase, which enables real-time localization.

The rest of the paper is arranged as follows. Section 2 presents the proposed system architecture and detail implementation of the genetic algorithm and machine learning models. Section 3 presents the experimental design. Section 4 presents the results and the discussion. The last section concludes the paper with future direction.

II. The Proposed Scheme

Similar to the architecture of an indoor Wi-fi Fingerprinting method, the system is divided into two-phase: the offline phase and the online phase. In addition, there is central storage which is responsible for saving collected data and trained machine learning models.

In the offline phase, it is divided into two modes: the initialization mode and the update mode. The initialization mode is used on the first run of the system when the storage is empty and it is illustrated in Fig. 1. First, a set of positions is designated for Wi-fi RSSI measurements, this set is denoted as Wi-fi Reference Points (WRP)s. Because the number of collected WRPs is small, their locations should cover the whole area. Then, for each WRP in the environment, the Wi-fi RSSI signals are carefully measured from all the access points and saved them to the central storage. The next step is letting users with a device equipped with Wi-fi transceiver modules and IMU sensors move around in the area. The raw motion data and Wi-fi RSSI are recorded and saved to the storage. When the system decided that it has collected enough user data, it would start analyzing and creating machine learning models. User raw motion data is separated from the user Wi-fi RSSI and they are sent to the PDR block to output processed motion data (step event, step length, and heading angle). Next, the WRP data and user processed motion data are forwarded into the block where a genetic algorithm and machine learning model is implemented to estimate user positions. Details of the implementation are presented in the later sections. Then, the estimated positions are used as the labels for the user Wi-fi RSSI data to form a new dataset. The new dataset is

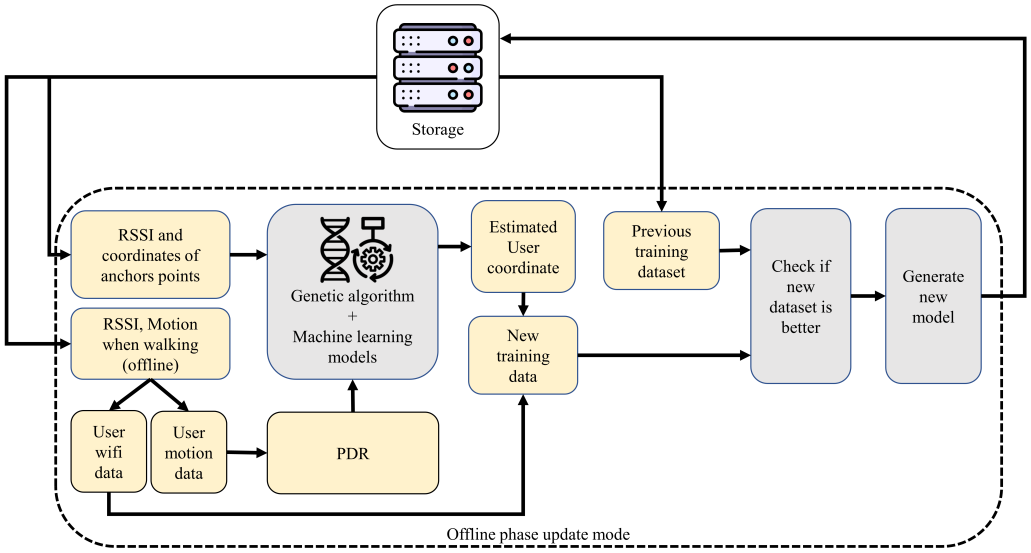


Fig. 2. Offline Phase Update Mode

combined with the WRP data to become the training dataset. Next, machine learning models were used to fit the training data then perform model tuning for better performance. The optimized model is later saved to the central storage.

The online phase is mentioned first for better chronological order. The online of the proposed system is used to estimate the position of the user using the trained model from the offline phase and it is illustrated in Fig. 1. First, user Wi-fi RSSI and motion data are recorded while the user moves. Then, the Wi-fi RSSI data are extracted from the user data and forwarded into the trained machine learning model to output user location. User data are saved to the central storage for further analysis in the offline phase.

The offline phase update mode is presented in Fig. 2. It is utilized when there are new user data collected during the online phase and system operator want to analyze the potential of this data to be used with the previous training dataset. First, the WRP data and user data are loaded from the storage. Then the user data is split into Wi-fi RSSI and raw motion data. The raw motion data is forwarded to PDR to extract processed motion data (step event, step length, and heading angle). Then, the WRP and processed motion data are sent to the genetic algorithm and machine learning block to create user positions similar to the initialization mode. Next, the previous training dataset is loaded from the central storage and compared with the new one. If the new dataset satisfies the metric, then it will be used as training data to generate a new model. Details of the metric are described in the next section.

A. Implementation of Genetic Algorithm and Machine Learning Block

The input of this block is WRP data and user motion data, the output is the estimated user position. WRP is defined as (1):

$$WRP_i = \{(RSSI_i^1, ..., RSSI_i^j, ..., RSSI_i^M), (x_i, y_i)\} \quad (1)$$

where $i = 0$ to N_{WRP} and N_{WRP} is the number of WRPs, M is the number of Wi-fi AP, $RSSI_i^j$ is the RSSI measurement from the j^{th} AP of i^{th} WRP, (x_i, y_i) is the coordinates of i^{th} WRP.

An assumption was made that a PDR algorithm was implemented and it processed the raw motion data from IMU sensors and output the step event, step length, and heading angle. Using the output, the distance and the angle are calculated between two consecutive Wi-fi RSSI measurement points. Then the user Wi-fi RSSI data and the processed motion data is in the form (2):

$$U_i = \{(RSSI_i^1, ..., RSSI_i^j, ..., RSSI_i^M), (\alpha_i, d_i)\} \quad (2)$$

Where $i = 0$ to N_U, N_U is the number of user data point, α_i is the angle formed by two vector $\overrightarrow{U_iU_{i+1}}$ and $\overrightarrow{U_{i+1}U_{i+2}}$ and d_i is euclidean distance between U_i and U_{i+1}

From the processed motion data $\{(\alpha_1, d_1), (\alpha_2, d_2), ...(\alpha_{N_U}, d_{N_U})\}$, the user route shape can easily be obtained. However, because of not knowing the user's starting point, it is not possible map the route to the environment. As the starting point can be any position in the area, if a searching algorithm is implemented to search for every possible location, it would cost a lot of time and computation resources. To tackle this, a genetic algorithm was proposed.

Genetic algorithm (GA) was first introduced by John Holland in 1960 in [24] and it has been applied as a method to solve optimization and search problems. GA can be divided into three main steps which are population initialization, fitness evaluation [25] and applying genetic operations. The algorithm repeats steps 2 and 3. For each repetition, a new generation is created. The solution is obtained and the algorithm stops when the fitness value has converged or GA has reached the maximum number of generations. Based on the structure of a standard GA, a version of GA was implemented to find the user position in the 2-dimensional coordinate system from

processed motion data. The following are the details of the implementation.

1) Population Initialization: First, the representation of an individual (chromosome) in the population is considered. Although searching for the user's starting point is the goal, the whole route also has to be considered because when calculating the error, it is important to calculate for the whole track. An array with each element consisting of two floating-point numbers representing the coordinates of the user in the environment is chosen and shown in (3). No encoding method method like Binary Encoding was used because it requires further computation for binary conversion and loss of precision.

$$p_i = \{(x_1, y_1); (x_2, y_2); ...(x_{N_U}, y_{N_U})\} \tag{3}$$

While randomly generating the initial population (first generation), it is important to determine the population size. The number varies in real cases because of factors like search space, processing capability, and environmental constraints. After trials, the population size is selected to be 100. Constraints were also put on the whole route that every point must be inside the range $[x_{min}, x_{max}], [y_{min}, y_{max}]$. This range is to prevent the generated track from not being too far off the indoor area. To get a random track, the starting point is randomly generated, then the rest of the track positions are calculated using processed motion data mentioned earlier.

Algorithm 1 Proposed Genetic Algorithm

INPUT: WRP data, user motion data
OUTPUT: Estimated user positions

begin
Set N_P, mutation rate ϵ_M, convergence condition E
$i = 0$
Initialize first population $P(i)$
Calculate fitness of $P(i)$
p = individual with highest fitness of $P(i)$
while E is not satisfied **do**
    Create empty $P(i+1)$
    Populate $P(i+1)$ using selection operator on $P(i)$
    Apply crossover operator on $P(i+1)$
    Apply mutation operator on $P(i+1)$
    Calculate fitness of $P(i+1)$ using ML and WRP
    $\widetilde{p}$ = individual with highest fitness of $P(i+1)$
    if (Fitness($\widetilde{p}$) > Fitness(p)) **then**
        $p = \widetilde{p}$
    end if
    Replace $P(i)$ by $P(i+1)$
    $i = i + 1$
end while
Output best p
end

2) Fitness Evaluation: The fitness value of an individual needs to show how well that individual perform compared to others. After the initial population is generated, a method to calculate the fitness of the user track by taking advantage of machine learning models was proposed. It is known that if a machine learning model was trained using accurate training data, then the error on the testing data would be small. Using this idea, the random track position is used as the label for the user Wi-fi RSSI data as the training set and the testing set is the WRP dataset. Let's say that the model prediction is (4).

$$\overline{WRP} = \{(\overline{x_1}, \overline{y_1}), (\overline{x_2}, \overline{y_2}), ..., (\overline{x_{N_U}}, \overline{y_{N_U}})\} \tag{4}$$

Then the positioning error between WRP and $\overline{WRP}$ is calculated using (5):

$$E(\overline{WRP}) = \frac{\sum_{i=1}^{N_{(WRP)}} \sqrt{(x_i - \overline{x_i})^2 + (y_i - \overline{y_i})^2}}{N_{(WRP)}} \tag{5}$$

The intuition that an individual which performs better would have a higher fitness value is better for comprehension. Therefore, the fitness value is computed using (6).

$$F(p_i) = \frac{1}{E(\overline{WRP_i})} \tag{6}$$

where $i = 1$ to N_p and N_p is the population size, $\overline{WRP_i}$ is the predicted WRP using the i^{th} individual in the population.

A simulation is carried out to prove that the fitness function in (6) can show how close and accurate the randomly generated track is compared to the real user track. Details are described in the experiment section.

3) Genetic Operations: Three operators that need to implement are selection, crossover, and mutation. For the selection operator, Roulette Selection was used. The probability of an individual being selected for the next generation is calculated using (7).

$$\mathbb{P}(p_i) = \frac{F(p_i)}{\sum_{j=1}^{N_p} F(p_j)} \tag{7}$$

The crossover operator was designed so that between two individuals, the one with higher fitness would have more contribution to the offspring. The amount of contribution is evaluated using the metric called fitness weight. If individual A $p_A = \{x_i^A, y_i^A, i \in [1, N_U]\}$ and individual B $p_B = \{x_i^B, y_i^B, i \in [1, N_U]\}$ are selected as parents, then the fitness weight of A and B is computed as in (8):

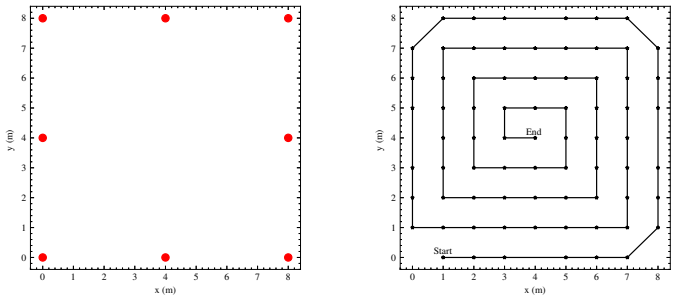
$$w_A = \frac{F(p_A)}{F(p_A) + F(p_B)}; w_B = \frac{F(p_B)}{F(p_A) + F(p_B)} \tag{8}$$

If the offspring of A and B is denoted as C $p_C = \{x_i^C, y_i^C, i \in [1, N_U]\}$ then the coordinates in C are calculated using (9), (10):

$$x_i^C = w_A * x_i^A + w_B * x_i^B \tag{9}$$

$$y_i^C = w_A * y_i^A + w_B * y_i^B \tag{10}$$

For mutation, mutations in the population are created by randomly created 2 values (dx, dy) which represent how far the whole track will be shifted in a 2-dimensional area. The mutation rate is chosen to be a small constant number because it helps the algorithm to converge faster. Details of the configuration are described in the experiment section. The steps of the genetic algorithm is shown in Algorithm 1.



(a) (b)

Fig. 3. (a)Wi-Fi Reference Points marked as Red (b) User Route where Black Marks Indicate Locations of RSSI Measurements

TABLE I. GENETIC ALGORITHM PARAMETERS

Parameter	Value
Population size	100
Maximum number of generation N_G	50
Coordinate boundary	$[-5, 15]$ meters
Mutation $[dx, dy]$ boundary	$[-3, 3]$ meters
Mutation rate	0.001
Convergence condition	Reaching N_G

III. EXPERIMENTAL DESIGN

The experiment was carried out in the laboratory which is a square area of 10 by 10 meters. The room has tables, chairs, computers, and other networking devices. The Wi-fi network is set up and 8 APs are placed around the room. First, The offline phase was performed by designating 8 points in the testing site as the WRP, their locations are illustrated in Fig. 3(a) Then, at each WRP, the RSSI measurements from 8 APs are collected and saved to the central storage. Because the processed motion data (heading angle and distance) were applied from other research and to ease the implementation of the experiment, the user motion data are simulated with random Gaussian noise (noise ranges from 5% to 30%). Then one person would follow the path of the simulated route. While going, the RSSI measurements are collected at marked points in Fig. 3b and saved the data to the central storage.

After acquiring all the necessary data, the next step is to run the genetic algorithm. In the population initialization step, the boundary constraints on generated coordinates are set to be in the range of $[-5, 15]$ (meter) in both axes. The constraints prevent the generated coordinates from being too far off and the negative range while creating a diverse population. Table I shows details of genetic algorithm parameters.

The fitness function needs predictions from the machine learning model to calculate fitness value. However, machine learning models usually take time to train and predict. As the result, it is not recommended to use too many models and perform hyperparameter tuning in this situation. Two machine learning models that have been widely used in indoor positioning research were selected which are Support Vector Machine Regression (SVR) and K-Nearest Neighbors Regressor (KNN). Fixed hyper-parameters were selected for both model in Table II and Table III with no tuning while calculating the fitness.

TABLE II. SVR MODEL PARAMETERS

Parameter	Value
Kernel	Radial basis function (RBF) kernel
Regularization parameter	1.0
Epsilon-SVR	0.1

TABLE III. KNN MODEL PARAMETERS

Parameter	Value
Number of neighbor	5
Weight function	Uniform
Algorithm	Select from Ball-Tree, KD-Tree and Brute force
Leaf size	30
Metric	Minkowski

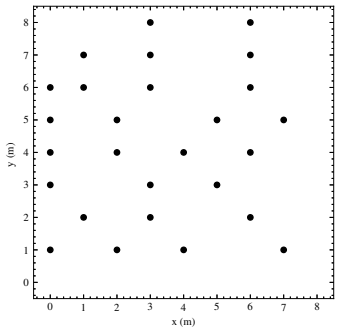


Fig. 4. Testing Point in the Experiment Area

When GA outputs the estimated user position, those positions are mapped to the user Wi-fi RSSI data and combined with the WRP to become training data. At this step, four machine learning models were selected: SVR, KNN, Multi-Layer Perceptron (MLP), and Random Forest (RF) to train and perform optimization on hyper-parameters to achieve the best possible results. The optimization method is Grid Search which was implemented in the sci-kit learn library.

Finally, the performance of the trained models is tested by collecting 27 random points with Wi-fi RSSI measurements as

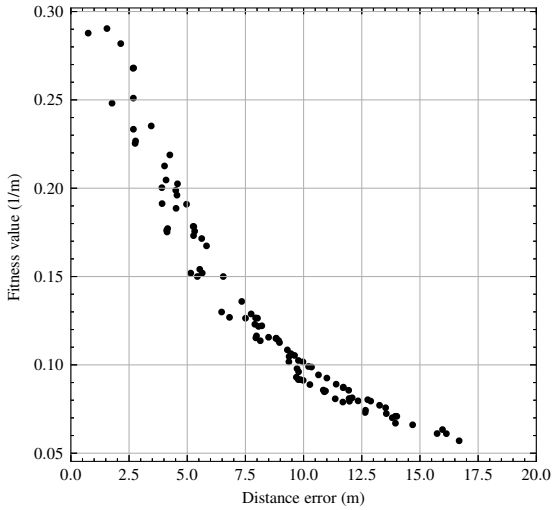


Fig. 5. Relationship between the Fitness Value and the Distance Error of 100 Random Tracks

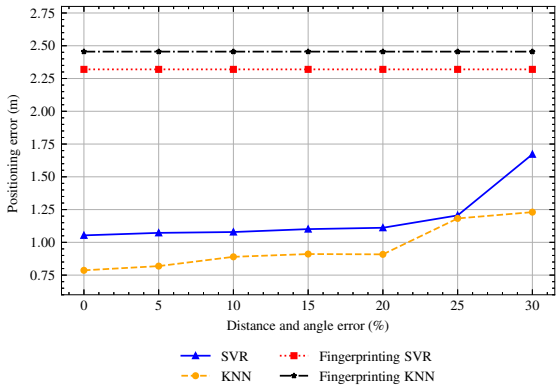


Fig. 6. Positioning Error of the Proposed Method and Conventional Fingeprinting using 8 WRPs with respect to Different Amount of Motion Error

illustrated in Fig. 4. Then the prediction error of each model is calculated and compared.

IV. Results and Discussion

A. Fitness Function Evaluation

In this part, the effectiveness of the fitness function on the collected data is illustrated. Using the processed motion data, the obtained route shape is similar to the one in Fig. 3(b). Then, 100 initial points in the environment were randomly generated. From those points, 100 user tracks were acquired. Equation (6) was used to calculate the fitness value of each track and (5) was used to measure the error of the randomly generated track to the real one. The relationship between the two values is shown in Fig. 5.

From Fig. 5 it is clear that when the fitness value is high, the distance error of the randomly generated track and the real track is low. This shows that the fitness function can create value that reflects how close a random track is to the real track. Looking closely at the top right of Fig. 5, it is noticeable that the track with the highest fitness value is not the one with the lowest positioning error. This shows that the fitness function can not find the absolute best because when the machine learning models were used to make a prediction, it is important to also account for the error in the WRP testing set. As it is impossible create an error-free dataset, the error is unavoidable and the only way to counter it is to carefully measure each point in the WRP dataset. Although the track with the lowest positioning error may not be found, the track with the highest fitness is guaranteed to be its neighbors, which gives a reasonable estimation.

B. Offline Phase Results

The positioning error of the estimated user track coordinates is computed using (5). The same machine learning models with the same configurations as in Table II and Table III were used in the conventional Fingerprinting method for comparison. In conventional Fingerprinting, ML models were trained on the WRP dataset, and then they were tested with the user Wi-fi RSSI dataset. Fig. 6 illustrated positioning error in relationship with different amounts of processed motion error (from 0 % to 30 %) of our proposed method and conventional Fingerprinting method.

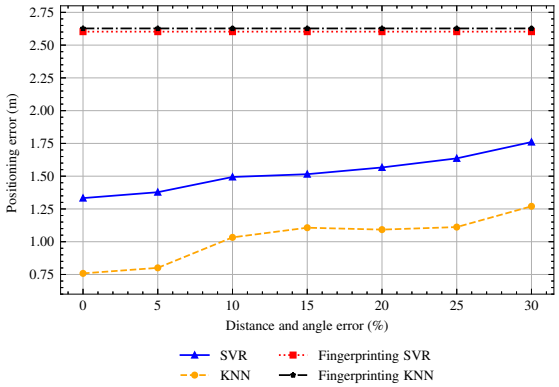


Fig. 7. Positioning Error of the Proposed Method and Conventional Fingeprinting using 6 WRPs with respect to Different Amount of Motion Error

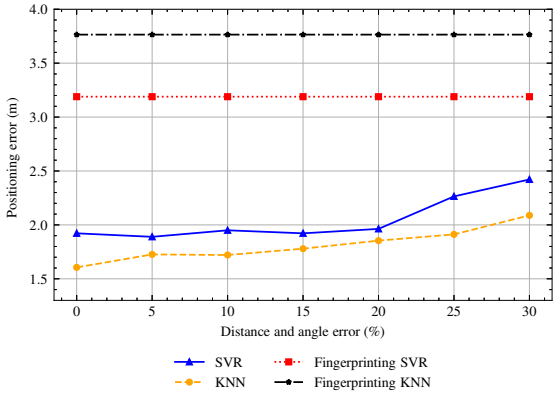


Fig. 8. Positioning Error of the Proposed Method and Conventional Fingeprinting using 4 WRPs with respect to Different Amount of Motion Error

From Fig. 6, the proposed genetic algorithm had better performance in both cases using the SVR and the KNN. With traditional Fingerprinting, it does not use motion data so the results are the same across the different amounts of motion error. For Fingerprinting the positioning error of the SVR is 2.32 meters and the KNN is 2.45 meters, which had been optimized by the Grid Search search algorithm. On the other hand, GA relies on motion data so when processed motion error rises, the positioning error of the proposed method also increases. From 0% to 20% motion error, the positioning error of the proposed method increased slightly from 1.05 meters to 1.11 meters for the SVR model, which is only 5.7% of the increased error. The same trend can be seen with the GA and KNN model where it rises from 0.78 meters to 0.9 meters. From 20% motion error onward, both GA models had a sharp rise especially for SVR at 30% motion error and KNN at 25% motion error. However, even at 30% motion error, the results of GA are 1.67 meters for SVR and 1.23 meters for KNN, which is still much better than the conventional Fingerprinting method. For comparison, our GA with the unoptimized SVR and KNN has positioning error 28% and 46% lower than that

of the Fingerprinting with the optimized models. However, it is worth noticing that although the SVR model is more complex and is supposed to have better performance than the KNN model, GA with KNN has better performance. As mentioned earlier no hyperparameters tuning was performed inside GA so both models may not be optimized and KNN initial parameter may be better than SVR.

To analyze the impact of the number of WRPs on the positioning error, the experiment was carried out with a different number of WRP. Fig. 7 and Fig. 8 show the results when applying the same method and configuration but with 6 and 4 Wi-fi Reference Points. When lowering the number of WRPs, it is expected that the conventional Fingerprinting method would have positioning error increased because of the smaller training dataset. The situation can be observed in both Fig. 7 and Fig. 8. The fingerprinting SVR at 6 WRPs has a positioning error of 2.60 meters, which increases by 12% compared to the similar one at 8 WRPs. In the case of Fingerprinting KNN, the positioning error is 2.62 meters, which is close to the Fingerprinting SVR. The proposed GA with SVR and KNN also depend on the WRP dataset to calculate fitness so the performance is also affected. In Fig. 7, GA with KNN has a lower positioning error than GA with SVR. Compared to the performance of GA with KNN using 8 WRPs, the one using 6 WRP is similar. On the other side, GA with SVR using 6 WRPs has positioning error significantly higher than the one using 8 WRPs. The difference between the two cases ranges from 0.3 to 0.4 meters.

Looking at Fig. 8, the difference in positioning error while using less WRP is even more noticeable. For all cases in proposed GA and Fingerprinting, they all experienced a sharp increase. Although GA with KNN using 4 WRPs still be able to maintain its position to be the best solution, the positioning error has seen rises ranging from 0.7 to 0.9 meters compared to positioning error of the same one using 6 WRPs, which makes the result become 1.6 meters at 0% motion error and reach up to 2.1 meters at 30 % motion error. GA with SVR using 4 WRPs comes behind KNN and the result is not too far off, it has a positioning error of 1.92 meters at 0 % motion error and gets up to 2.42 meters at 30 % motion error. In the case of conventional Fingerprinting, KNN has the worst result with a positioning error of 3.76 meters, which is 35 % higher than the same version that uses 6 WRPs. Fingerprinting with SVR using 4 WRPs achieves a result at 3.19 meters, which is better than the one with KNN but poor compared to the proposed GA with both models.

From the above observations of the conducted experiment, the proposed GA achieves better results than the conventional Fingerprinting approach. However, as the proposed GA uses user-processed motion data to create labels for the user Wi-fi RSSI data, motion error can greatly affect the estimated labels. Another factor that also has a big impact on the proposed GA performance is the number of collected WRPs. Although using fewer WRPs means the degradation of the results, the experiment had demonstrated that even with just 8 WRPs in a 10 by 10 meters area, the proposed GA was able to produce accurate estimations.

After GA outputs the estimated user track position, the track generated by the KNN model was selected because it has the lowest positioning error. Then, mapping from the estimated positions to user Wi-fi RSSI and combining with WRP dataset was done to create training data for 4 machine learning models: SVR, KNN, MLP, and RF. After the training and model optimization process, all models were saved to the central storage for evaluation in the online phase.

C. Online Phase Results

After loading the trained model from the central storage, 27 points with RSSI measurements illustrated in Fig. 4 were used as the testing dataset for evaluating model performance. For comparison, Fingerprinting method was employed, 4 ML models similar to the last training step in the offline phase were used and they were trained and optimized on the WRP dataset. Positioning error (5) was still used as the metric to measure the distance between the estimated points and the ground truth. In addition, positioning error was calculated at different amounts of motion data in the training data from the previous section. Results of the proposed SVR, KNN, MLP, and RF along with its Fingerprinting version were shown in Fig. 9a, Fig. 9b, Fig. 9c, and Fig. 9d respectively.

Looking at the 4 graphs in Fig. 9, it is clear that all 4 models which were trained using the estimated positions achieved lower positioning error than the Fingerprinting version. It can be seen that motion error of the training set reflects on the performance of the model on the testing set and they form a linear relationship. Across 4 models, although there were some exceptions, the general trend is when the motion error of the training set grows, the positioning error of the trained model in the testing dataset increases.

In Fig. 9(a) the positioning error of the proposed SVR model on the testing set was 1.65 meters at 0% motion error and it reached 1.89 meters at 30% motion error. Conventional Fingerprinting using the SVR model had a positioning error of 2.24 meters across all levels of motion error. The difference between the two models in the worst case is 0.35 meters, which makes the Fingerprinting SVR positioning error 18% higher than the proposed SVR model. In Fig .9b, the result of the Fingerprinting KNN is 2.23 meters which is close to the Fingerprinting SVR. The proposed KNN positioning error gradually increased from 1.67 meters at 0% motion error to 1.76 meters at 25% motion error. Then a sudden jump at 30% motion error happened, which made the positioning error of the proposed KNN become 2.02 meters. Results of MLP models were illustrated in Fig. 9c. The proposed MLP error at 0% was 1.75 meters, which was the highest among the 4 proposed models but from 5% motion error, the performance was close to the proposed SVR and KNN. Finally, Fig. 9d illustrates the results of the RF model. It can be seen that the positioning error of the Fingerprinting RF was 2.3 meters and it was the highest across all models and methods used in this section. While the fingerprinting RF result was poor, the proposed RF achieved good results with positioning error at 0% motion error was 1.63 meters and got up to 1.93 at 30% motion error.

Similar to the previous section, the impact of the number of WRPs on the proposed model performance is analyzed. Training data with the lowest positioning error using 6 WRPs and 4 WRPs were selected from the previous section and became the training data for 4 machine learning models. Fig. 10 shows the comparison of positioning error among 4

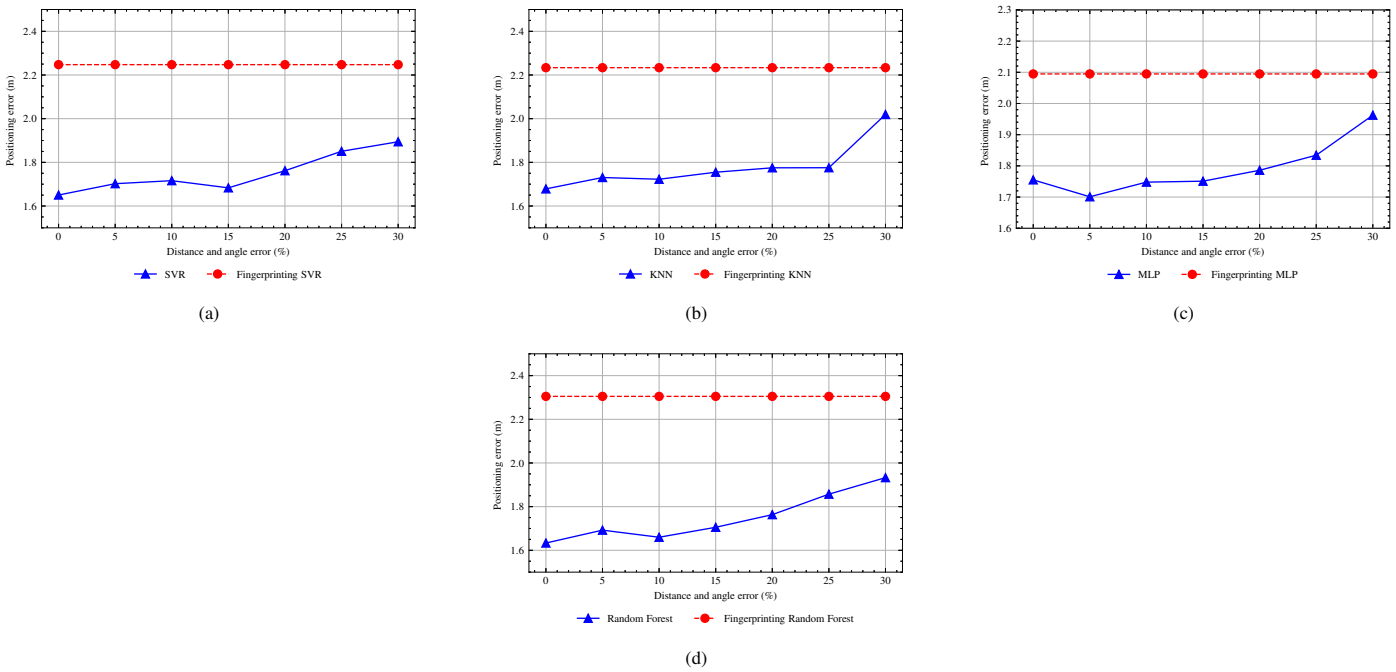


Fig. 9. Positioning Error of 4 Machine Learning Models using the Proposed Method and Conventional Fingeprinting with respect to Different Amount of Motion Error. (a) SVR, (b) KNN, (c) MLP, (d) RF

proposed models with different numbers of WRPs across 6 levels of motion error.

From Fig. 10 it can be seen that using fewer WRPs would result in the increase of positioning error for all models. However, the difference between using 8 and 6 WRPs is not significant while there is a huge gap between using 8 WRPs and 4 WRPs.

At 5% motion error in Fig. 10(a), the positioning error of all models that use 8 WRPs are close, ranging from 1.69 meters (RF) to 1.73 meters (KNN). When using 6 WRPs, all model's positioning errors had slight increases. The proposed KNN model still had the worst result at 1.86 meters following are MLP, SVR, and RF. At 6 WRPs, a big jump of positioning error can be observed across all 4 proposed models. This time, RF had the most significant increase and peaks at 2.45 meters, which was also the highest positioning error among others.

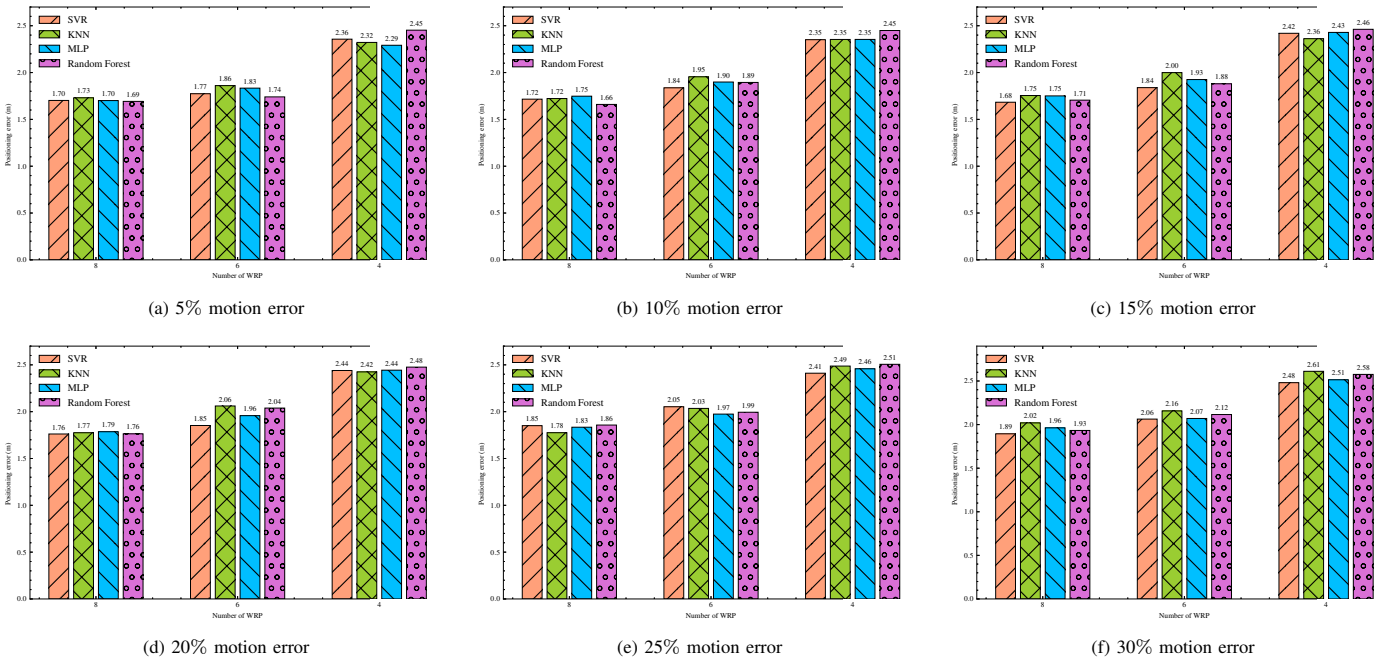
As can be seen, the trend in Fig. 10(b) to Fig. 10(c) is similar to Fig. 10(a), especially with the proposed RF model. While its results among the lowest positioning error at 8 WRPs and 6 WRPs, the performance dramatically decreased at 4 WRPs. Other proposed models had close positioning error and they only had a mild rise across levels of motion error. The proposed models at 10 % motion error that use 6 WRPs had errors ranging from 1.84 meters (SVR) to 1.95 meters (KNN) and they reached the range from 1.85 meters to 2.06 meters at 20 % motion error. A similar case can be observed with models that used 4 WRPs at 10 % motion error had positioning error ranging from 2.35 meters to 2.45 meters and they jumped to the range from 2.42 meters (KNN) to 2.48 meters (RF). As the motion error gradually increased and peaked at 30 %, Fig. 10(f) shows the worst case for all models. From Fig. 10(f), it can be seen that SVR has the best performance among others across different numbers of WRPs. It has a positioning error of 1.89 meters at 8 WRPs and peaks at 2.61 meters at 4 WRPs. On the other hand, KNN has the worst performance with positioning error at 2.02 meters at 8 WRPs and got up to 2.61 meters at 4 WRPs.

From the above analysis, it is clear that the number of WRP and proposed model performance has a close relationship. When less number WRP were used, the positioning error of the estimated user positions will rise. In addition, the training data would have less accurate data points than before. For those reasons, it is expected that the trained model performance would get poorer when the number of WRP drops. In real cases, it is important to select an appropriate number of WRP to collect. When system users want to have more accurate results, then more WRP may need to be collected. Another recommendation is the designation of the WRPs location should form a grid that covers the whole area so that they can capture the pattern of Wi-fi signal in every position.

The above experiment presented the online phase in the first run of the system. In practice, both user motion data and Wi-fi RSSI are collected and analyzed for potential usage. If the positioning error of a new user data is lower than the one that was used previously, then the model can be retrained for better performance. This creates a close loop system where user data are collected and models are updated continuously to suit the indoor environment.

V. Conclusion

This paper proposed a new architecture in the offline phase of the Fingerprinting method. Our proposed architecture takes advantage of the user motion data and GA to create labels for the user's Wi-fi RSSI data. It enables our models to have



(a) 5% motion error (b) 10% motion error (c) 15% motion error

(d) 20% motion error (e) 25% motion error (f) 30% motion error

Fig. 10. Positioning Error of 4 Machine Learning Models with respect to the Different Number of WRPs Across Different Amount of Motion Error

more training data to accurately predict user location while reducing the amount of data that needs to be collected in the conventional Fingerprinting method. However, it is worth mentioning that our proposed offline phase procedure requires heavy computation as the fitness value of every individual in the population of GA needs to perform machine learning model training and prediction. In addition, the proposed method relies on motion data so any error, noise, and interference in the motion data can affect the output of GA and machine learning block. In the future, more studies on the application of the genetic algorithm and also the field of evolutionary computation will be conducted to improve the performance and reduce the computational complexity of the existing system.

REFERENCES

[1] S. A. Golden and S. S. Bateman, "Sensor Measurements for Wi-Fi Location with Emphasis on Time-of-Arrival Ranging," *IEEE Trans. on Mobile Comput.*, vol. 6, no. 10, pp. 1185–1198, Oct. 2007. [Online]. Available: http://ieeexplore.ieee.org/document/4294899/

[2] C. Yang and H.-r. Shao, "WiFi-based indoor positioning," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 150–157, Mar. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7060497/

[3] F. Subhan, H. Hasbullah, A. Rozyyev, and S. T. Bakhsh, "Indoor positioning in Bluetooth networks using fingerprinting and lateration approach," in *2011 International Conference on Information Science and Applications*. Jeju Island: IEEE, Apr. 2011, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/document/5772436/

[4] D. Ahmetovic, M. Murata, C. Gleason, E. Brady, H. Takagi, K. Kitani, and C. Asakawa, "Achieving Practical and Accurate Indoor Navigation for People with Visual Impairments," in *Proceedings of the 14th International Web for All Conference*. Perth Western Australia Australia: ACM, Apr. 2017, pp. 1–10. [Online]. Available: https://dl.acm.org/doi/10.1145/3058555.3058560

[5] C. Jihong, "Patient Positioning System in Hospital Based on Zigbee," in *2011 International Conference on Intelligent Computation and Bio-Medical Instrumentation*. Wuhan, China: IEEE, Dec. 2011, pp. 159–162. [Online]. Available: http://ieeexplore.ieee.org/document/6131776/

[6] H. Kobayashi and A. F. Molisch, "Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks," *IEEE Signal Processing Magazine*, vol. 22, no. 4, p. 15, Jul. 2005.

[7] M. N. Husen and S. Lee, "Indoor human localization with orientation using WiFi fingerprinting," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication - ICUIMC '14*. Siem Reap, Cambodia: ACM Press, 2014, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2557977.2557980

[8] S. He and S.-H. G. Chan, "Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 466–490, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7174948/

[9] F. Zafari, A. Gkelias, and K. K. Leung, "A Survey of Indoor Localization Systems and Technologies," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8692423/

[10] D. Tinh Pham and T. T. Ngoc Mai, "Ensemble learning model for Wifi indoor positioning systems," *IJ-AI*, vol. 10, no. 1, p. 200, Mar. 2021. [Online]. Available: http://ijai.iaescore.com/index.php/IJAI/article/view/20603

[11] P. Bahl and V. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, vol. 2. Tel Aviv, Israel: IEEE, 2000, pp. 775–784. [Online]. Available: http://ieeexplore.ieee.org/document/832252/

[12] A. Anjum and M. U. Ilyas, "Activity recognition using smartphone sensors," in *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*. Las Vegas, NV: IEEE, Jan. 2013, pp. 914–919. [Online]. Available: http://ieeexplore.ieee.org/document/6488584/

[13] W. Kang and Y. Han, "SmartPDR: Smartphone-Based Pedestrian Dead Reckoning for Indoor Localization," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2906–2916, May 2015. [Online]. Available: http://ieeexplore.ieee.org/document/6987239/

[14] A. Jimenez, F. Seco, C. Prieto, and J. Guevara, "A comparison of Pedestrian Dead-Reckoning algorithms using a low-cost MEMS IMU," in *2009 IEEE International Symposium on Intelligent Signal Processing*. Budapest, Hungary: IEEE, Aug. 2009, pp. 37–42. [Online]. Available: http://ieeexplore.ieee.org/document/5286542/

[15] A. Brajdic and R. Harle, "Walk detection and step counting on unconstrained smartphones," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. Zurich Switzerland: ACM, Sep. 2013, pp. 225–234. [Online]. Available: https://dl.acm.org/doi/10.1145/2493432.2493449

[16] Q. Wang, L. Ye, H. Luo, A. Men, F. Zhao, and Y. Huang, "Pedestrian Stride-Length Estimation Based on LSTM and Denoising Autoencoders," *Sensors*, vol. 19, no. 4, p. 840, Feb. 2019. [Online]. Available: http://www.mdpi.com/1424-8220/19/4/840

[17] M. J. Abadi, L. Luceri, M. Hassan, C. T. Chou, and M. Nicoli, "A collaborative approach to heading estimation for smartphone-based PDR indoor localisation," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. Busan, South Korea: IEEE, Oct. 2014, pp. 554–563. [Online]. Available: http://ieeexplore.ieee.org/document/7275528/

[18] S. J. Julier and J. K. Uhlmann, "A New Extension of the Kalman Filter to Nonlinear," *Proc. SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition VI*, p. 12, Jul. 1997.

[19] S. O. H. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," 2010.

[20] A. Khalajmehrabadi, N. Gatsis, and D. Akopian, "Modern WLAN Fingerprinting Indoor Positioning Methods and Deployment Challenges," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 3, pp. 1974–2002, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7874080/

[21] B. Ferris, "WiFi-SLAM Using Gaussian Process Latent Variable Models," *Proceedings of the 20th international joint conference on Artifical intelligence*, Jan. 2007.

[22] A. Naguib, P. Pakzad, R. Palanki, S. Poduri, and Y. Chen, "Scalable and accurate indoor positioning on mobile devices," in *International Conference on Indoor Positioning and Indoor Navigation*. Montbeliard, France: IEEE, Oct. 2013, pp. 1–10. [Online]. Available: http://ieeexplore.ieee.org/document/6817856/

[23] I. Constandache, R. R. Choudhury, and I. Rhee, "Towards Mobile Phone Localization without War-Driving," in *2010 Proceedings IEEE INFOCOM*. San Diego, CA, USA: IEEE, Mar. 2010, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/document/5462058/

[24] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, 1st ed., ser. Complex adaptive systems. Cambridge, Mass: MIT Press, 1992.

[25] A. L. Nelson, G. J. Barlow, and L. Doitsidis, "Fitness functions in evolutionary robotics: A survey and analysis," *Robotics and Autonomous Systems*, vol. 57, no. 4, pp. 345–370, Apr. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0921889008001450

Geolocation Mobile Application to Create New Routes for Cyclists

Jesús F. Lalupú Aguirre
University of Sciences and Humanities
Faculty of Sciences and Engineering
Lima Peru

Laberiano Andrade-Arenas
University of Sciences and Humanities
Faculty of Sciences and Engineering
Lima Peru

***Abstract*—In Peru in recent decades it has undergone unexpected changes, often generating chaos among the population, such as the excess of vehicles that travel daily on the roads generating pollution. This has led people to seek alternatives, such as the use of bike, as a means of transportation. The objective is develop a mobile application for the creation of alternative routes for cyclists. For them we have carried out a survey of 50 people dedicated to the field of cycling as well as people who do not exercise it in order to collect data, analyze it and create mechanisms that help these users. This application was developed in Android Studio implementing free libraries to achieve its geolocation in a way that provides all the facilities for the cyclist to move. For the process of creating this application, the Scrum methodology was used, the design of the prototype is done in Adobe Photoshop. It was obtained as results of the investigation carried out in the survey that 75% of the people are satisfied with the use of the application, 60% responded defining it as very good and 100% answered yes they would recommend the application.The investigation is of importance, since it would allow as future work the reduction of environmental contamination.**

***Keywords*—Android studio; cyclists; mobile application; geolocation; scrum**

I. Introduction

During the years the vehicular influx has collapsed in the Peru is faced with a large-scale social problem, of a critical and unsustainable nature [1]. In addition to this, the crisis would saturate as a result of the covid-2019 that paralyzed the whole world submerging people to distance themselves between families, friends and society [2]. This has motivated many people who usually made use of public transport and in some cases private transport seek other alternatives with which they can move to their jobs, carry out their recreational activities among other.

Cyclists in turn have merged their transport tool and have originated the cyclotourism activity that has been developing rapidly in recent months, and even so there is the possibility of considerable growth in the coming months, which has been reflected in the purpose of purchase and search of bicycles by customers [3]. This has generated that friends, athletes, users get together to visit tourist places in the capital as well as recreational spaces located in different parts of the city for recreational or sports purposes. That somehow seek to spread this activity that in some way is an alternative to city vehicular chaos [4].

In the development of this research, a mobile application oriented to the use of this minor means of transport has been created, a mechanism that help people who have a mobile device and a bicycle to move around the city without the fear that many times invades us like the fear of buses, taxis, trucks or any high-speed vehicle including linear motorcycles, traveling on vehicular roads due to lack of bicycle lanes, the fear of being run over, also nerves when driving and robberies.

The main reason for this research development is based on the fundamental problems generated by the lack of bicycle infrastructure, the lack of information that users have to face in a complicated scenario due to the lack of bicycle paths, dangerous crossings, the lack of signage and obstacles. during the tour.

Consequently, the research carried out on this problem on the bicycle as a means of transport was the result of the increase of this means in recent months. As well as the interesting ideas that have for carrying out the application, giving it added value in its basic functions in which it will perform, such as the search for bicycle lanes in the city, the location on the map and navigation during the transfer from point A to point B.

For the development of this research, the Scrum methodology was used as a framework [5] [6], which was used as a reference for the development of this project, with the main objective of implementing a mobile application in order to create new cycling routes in metropolitan Lima. Research contributes to solving problems such as vehicular chaos for some users, contributing to tourism in the city, increasing sales of stores dedicated to this area.

The article is made up of the following sections: in Section II, the review of the literature where the different investigations are analyzed; in Section III, the methodology was used, which allowed sequencing the steps to be carried out; in Section IV, results and discussions of the investigation; and the section V conclusions; the Section VI future work.

II. Literature Review

The following project aims to generate guidelines and indications for the development of a mobile application implemented with the Global Positioning System (GPS) oriented to cyclists in the city of metropolitan Lima, Due to the high demand for bicycles and skateboards / scooters in the city as a result of the excess of automotive transport, also with the use of smartphones and the GPS navigation system that is a very helpful tool that for the trip is more efficient and agile,

informing them about the routes predetermined by the search engine, it also shows the alternate routes from the city, the main points of interest and tourist attractions.

According to the authors [7], Law N ° 30936 decreed by the Peruvian state and the actions taken by the municipality of metropolitan Lima refer that this law is fundamental for the project, as it reflects the fundamental axis on the evolution of the bicycle as a means of transport, providing improvements so that the bicycle lane network, committing itself to the growth of this network in the short and long term. The author [8] used the Scrum methodology for the development of your project, using the Android operating system because it has 72.6% of the Peruvian market, important data for the research.

Meanwhile the author [9], it refers that in Peru, given the existence of the need of the unsatisfied market to which it is directed, the project of mobile applications for cyclists is a satisfactorily valuable tool and that also the bicycle as a means of transport is an issue to be deepened due to the current situation of the country.

What's more [10], indicates that this social problem should not only be treated like any innovation project, but also from the legal framework starting from the literature review because the analytical-synthetic method is used in addition to employing a qualitative analysis with the use of interviews to expert to complement the research and get a better focus. As well as [11], concludes that in addition to being an innovation tool to be used, the objective will be to carry out educational campaigns and workshops on the advantages and disadvantages of using bicycles as a means of transport focused mainly on schools and universities, as well as work centers in order to better project information and not minimize the risks of insecurity.

For the author [12], mentions that during the last decade mobile devices the world has become an important technological tool for society so much so that it has been transformed from a normal computer to something so small that can carry it in the pocket which at the same time allows us to perform various operations. In addition to contributing to the development of society, in an educational and cultural way.

In the research carried out by the aforementioned authors, they determine the importance of the implementation of this application as a facilitating means for cyclists dedicated to sports, businesses or any natural person who sees in the bicycle an alternative of urban transport. During the research processes, some of the authors mentioned did not take the problem on a large scale as the problem of environmental pollution that could decrease with the use of bicycles, but instead focused on the increase in tourism as a benefit for the private sectors dedicated to to the bicycle industry and not as a good for the entire Peruvian society. That is why the research work was carried out focusing on being able to use the mobile application to be able to create alternative routes for itself to contribute in giving improvements to the society as pollution reduction and reduce excess traffic.

III. Methodology

For the research work, the Scrum methodology was applied as it is an agile methodology that allows us to work the processes incrementally generated by the phases and roles that was detailed according to the meetings agreed in the process. However, the modifications that are presented can be done at any time since this framework allows us to do it both in the final part as well as in any of the deliverables or iterations as shown in Fig. 1 [13].

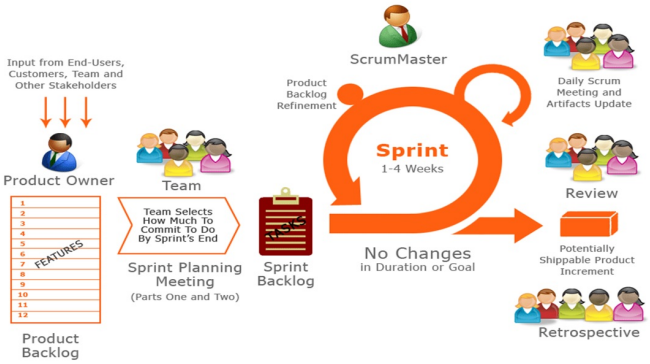


Fig. 1. Show Scrum Methodology [14].

A. SCRUM Phases.

They are the steps to follow for the development of a project. According to this framework, they are defined by times that guarantee the implementation of our mobile application and are as follows, as shown in Fig. 2.

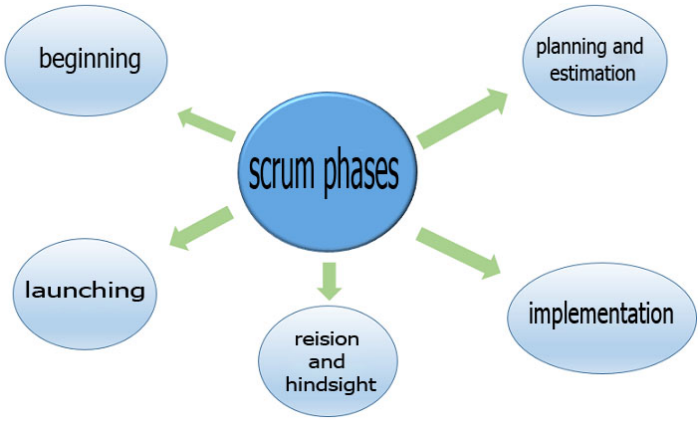


Fig. 2. Show the 5 Phases of Scrums

1) Beginning: It is where the research study begins, where the project is analyzed looking for the fundamental needs for each spring, which are the deliverables that was developed during the development of the project, in addition to asking questions such as: What do I want? How do I want? and When do I want? The vision and mission of the project must be created, identifying the Scrum master who was the leader of the group that has the necessary characteristics of a true leader.

2) Planning and Estimating: This stage is where it is created, in addition the user stories are identified, the user's requirements are studied, the iterations in which the mobile application was developed are created, considering that this

is the most important phase of the project, where the master Scrum delegates the roles of each participant in the project.

3) Implementation: is considered the stage of greatest distress not only of skill but requires mental work, since the defined project begins to be created, in addition to reflecting the ideas given in the meetings where it was arranged how to optimize the work, in this stage the deliverables, as well as the necessary tests are carried out taking into account that at this stage no modifications should be made since this is seen in the planning stage, but because this methodology facilitates the developer to make the modifications in any of the iterations, could do it if it were the case and what would help improve the deliverables and therefore the project in general.

4) Review and Retrospective: In this stage, all the spring or deliverables are validated, where the work team has to carry out the necessary self-criticisms in addition to making the corresponding improvements to achieve the favorable scope of the work [15].

5) Launching: This is the last phase of the work, it is where the finished product has to be delivered to the client, achieving the best expectations, in addition to the acceptance of the end user.

B. System Requirements

These are all the requirements that allowed us to create the mobile application based on the information obtained to create the user requirements, having these requirements, the iterations were created.

Table 1 shows the user stories, which were raised according to the data collection, as can be seen, each story has an acceptance criterion which allows determining to what extent the advance of the mobile application is acceptable, this it was fulfilled by the work team, it also allows to reference the working mechanism behind so that the user can be satisfied. In addition, each user story has its description of the behavior of the application and its functionality.

C. Product Creation Backlog

This section is made up of a list of user stories or the requirements raised by the user that are ordered according to their priority in which they are estimated to become iterations.

Table II shown the iterations carried out during the application creation process. Accordingly, the iterations and deliverables will be analyzed.

TABLE I. ESTIMATION OF USER STORIES.

User story	Description	criteria of acceptance
Register User.	As a user I need to create an account to be able to use the application with my username and password.	Register by mail or Google account and be my username and password.
Login.	As a user, I need to authenticate my account using my username and password to be able to use the application.	Login by email or Google account.
Show my location on the Map.	As a user I need to self-authenticate my account using my username and password to be able to use the application.	Login via email or Google account.
Show current city bike lanes.	As a user I need to be able to view the city bike lane.	It allows you to view the city bike lane network.
Search Address.	As a user I need to be able to view the address I am looking for to be able to move.	Search through the address drawn or by places where I want to go.
Generate default route or select route on the map.	As a user i need to be able to visualize the route that will be generated to go safely.	Generate the routes through the bicycle lanes without using highly dangerous roads and show the approximate time that my trip will last.
Navigate during the route.	As a user I need to be able to navigate during the route through audio directions.	Show the arrival time on the screen, make directions through the default Google voice.
Create Favorite Destinations.	As a user I need to be able to create, edit and delete favorite places, this way to allow quick access without the need to search for it.	Show the list of favorite places.
Register places of interest or tourist.	As a user I need to be able to view the address I am looking for to move and to be able to register it.	Search through the back address or by places where I want to go.
Create custom routes and be able to save them.	As a user I need to be able to create custom routes and use them in future trips.	The routes are assigned to the menu create route, list the routes.
Show notifications.	As a user I need to be able to view notifications related to the application.	notifications must be used.
Application options menu.	As a user I need to be able to view all the menu options.	access by sliding to the main menu.
Create shared route.	As a user I need to be able to share my routes or trips that I take as family, friends, as well as other users to be able to access shared routes.	It must be shown in the routes section created by the user.
Forgot password.	As a user I need to be able to reset my password in case I forget.	To reset the password the app will send a validation message to the initially registered email.

TABLE II. BACKLOG CREATION

User story	Estimate	Sprint	Estimation by sprint
Show my Location on the map.	3		
Show current bike lanes.	5		
Search direction.	5	1	34
Generate default route or select from map.	5		
Navigate along the route.	13		
Register user.	5		
Login.	5	2	28
Forget password.	5		
Register places of interest or tourist.	5		
Register Trips.	7		
Show notifications.	6	3	21
Create Favorite destinations.	5		
Create custom routes and save them.	9		
Application options menu.	9	4	29
Create shared routes.	7		

D. Application Planning

As shown in Fig. 3, the Scrum phases that were implemented in the research were carried out, this helped to have more order in the work carried out and meeting the objective.

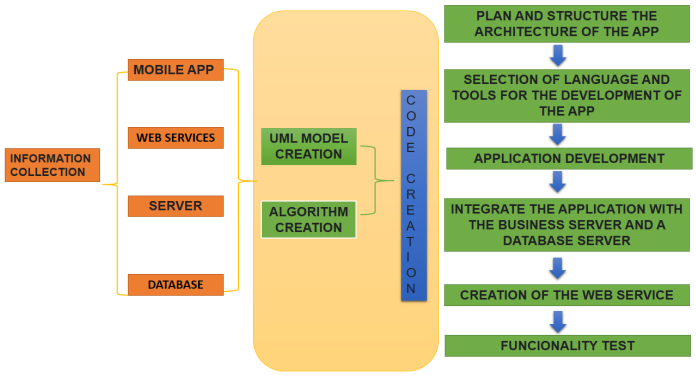


Fig. 3. Flow Chart using Scrum Methodology.

1) Planning and Structure of the Mobile Application: It is where the architecture of the work process for the implementation of the mobile application is defined, according to the data collected in the research references by the aforementioned authors and according to the data of the survey carried out.

2) Language Selection and Development Tools: At this stage, with the help of previous research, was seek to implement our new work mechanism, tools that are facilitators for the work to be carried out.

3) Application Development: According to what has been proposed, the developers will have established times for each iteration according to what was proposed in the meetings, in addition the planned times are from 2 to 8 days per iteration, during this process the mobile application is created progressively.

4) Integration of the App with the Business Server and the Database Server: At this stage the developers was the task of implementing the corresponding connections for their application functionality.

5) Creation of the Web Services: At this point the developers will start the communication process between the application with the database.

6) Tests: At the end of each iteration corresponding to each deliverable, the necessary tests were carried out for its compliance.

E. Data Collection

For the following investigation, 5 questions were formulated to around 50 people on the public road where some cyclists usually travel. In some cases it is possible to survey passers-by who do not travel by bicycle but who were interested in an application that helps them to use this means of transport. In addition, this helped to collect data and discover the interests of users or people who would suck in the future through this medium.The instrument was validated by the judgment of experts in the field; giving for approved the validation of the instrument.

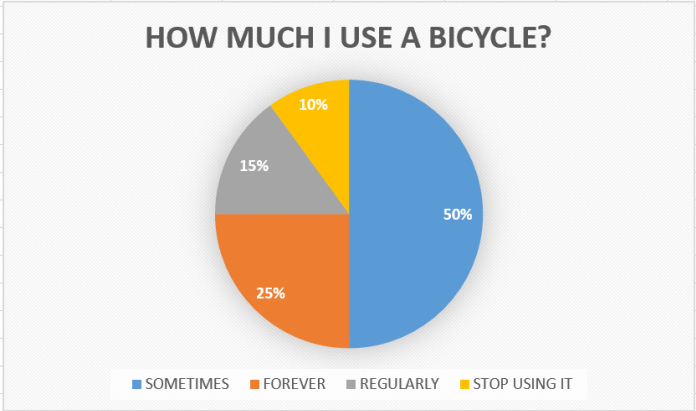


Fig. 4. Show Survey Results

Fig. 4 shows the statistical data of question number 1 in which the following results were obtained:

How much do you use a bicycle?

the responses were as follows:

- The equivalent of 50% of those surveyed answered that they sometimes use this medium.
- 25% of those surveyed answered that they always use. bicycle to move everywhere.
- 15% of those surveyed do it on a regular basis.
- 4% responded that they stopped using.

F. Mobile Application Architecture

For the design, the work team had to be previously assembled, likewise to propose the structure that the mobile application will have, carry out simple prototypes, define the services to be implemented, how it will work, in how many modules they will be carried out.

G. Structure of the Application

Fig. 5 shows the behavior of the application, where the final beneficiary is the cyclist, this had a web server, a database to store them, a messaging server, it will work with the satellite geolocator implementing free libraries, To use this tool, you must have a mid-range mobile device for use in addition to basic requirements.

H. Tools for the Implementation of the App

For the development of this application, different tools were used such as Android Studio, which is the platform where the application was encoded, the free libraries provided by Google maps were also used as geolocation tools, it also has a web service and a database data as detailed below.

1) Android studio
It is a programming software for the Android platform that contains all the necessary and explicit tools

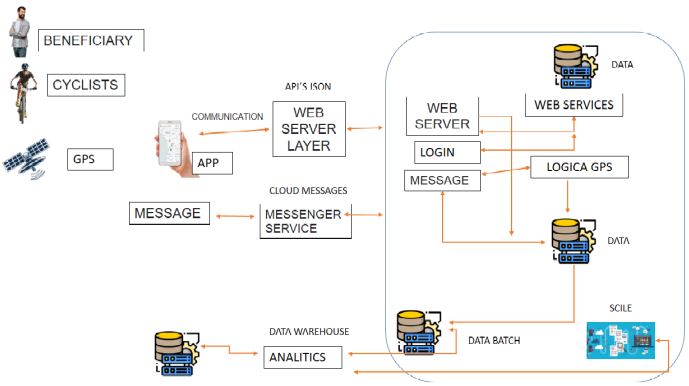


Fig. 5. Architecture of the Mobile App.

for the development of applications in this operating system (Android), it allows to carry out a number of configurations as well as import libraries.

2) Web services
It is an information system that exchanges XML-type messages with different systems that use different protocols such as HTTP [16], allowing communication between them.

3) Geolocation
precise location of the equipment or device, locate it directly on the search map from the satellite and bring it closer to the places closest to it [17], in order to provide the best service for it. be connected to the internet service for its operation to be able to process the chains in its interaction with the user regardless of the distance in which they are.

4) Database
The database system is in charge of managing the data generated in the application [18], which is used when necessary according to the will due to its data management capacity.

5) JSON
It is a simple data format that is used in programming for data exchange, this format is based on the JavaScript language, however it is very familiar to programmers because it uses different programming languages as a whole, facilitating their understanding [19].

6) Firebase
It is a type of database that allows synchronizing data in real time by storing it in the cloud, unlike other firebase, it does not use tables, much less records. But rather converts them into a JSON format with a password for each user at the same time, allows you to update the Data without the need to enter codes, they are simply saved or updated with the simple fact of being connected to the server [20].

7) OpenStreetMap
It is a free and editable map all over the world created by volunteer programmers with a free license with easy access to both the images of the map that at the same time can be related to devices that have GPS [11].

8) Photoshop
It is a graphic design software dedicated to the graphic industry due to its high quality of tools to execute all the editing processes that allow graphic editors to carry out more professional work.

IV. Results and Discussions

A. Application Satisfaction

According to the data obtained in the first survey that was carried out at the beginning of the project, the mobile application was implemented in order to solve the problems that the users expressed in their answers, as final results, the vast majority of positive responses were obtained in which gives the viability of this application. 5 questions were formulated to 50 people in order to test the application and how satisfactory its use is for end users, the results were the following:

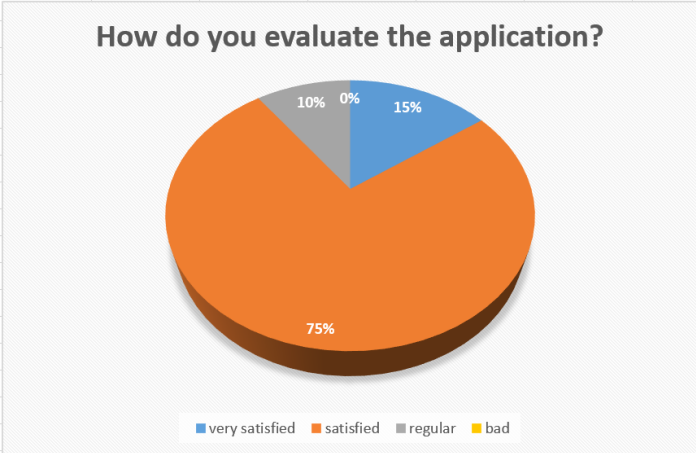


Fig. 6. Shows the Level of Satisfaction of the Application.

Fig. 6 shows the result of the survey carried out with 50 people on how satisfactory they found using the application, assigning them ratings such as: very satisfied, satisfied, fair, bad; getting a positive result as shown:

- 15% of people who are very satisfied.
- 75% of people who were satisfied with the application.
- The 10% that seemed regular to the application.
- A negative 0%.

Fig. 7 shows the results of question number 2 in which the respondent was asked whether the experience of riding a bicycle using the application improved. The results were the following:

- 50% answered what is experience improved
- 30% responded that their experience was fair.
- 15% answered that they only improved a little more than usual.
- 5% answered that nothing improved.

Fig. 8 shows the results of question number 3 in which the respondent was asked, How does he familiarize himself with the application? The results were the following:

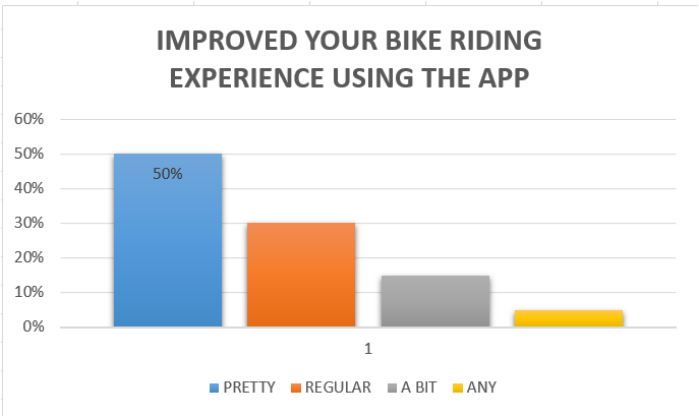


Fig. 7. Show the Results of Question 2.

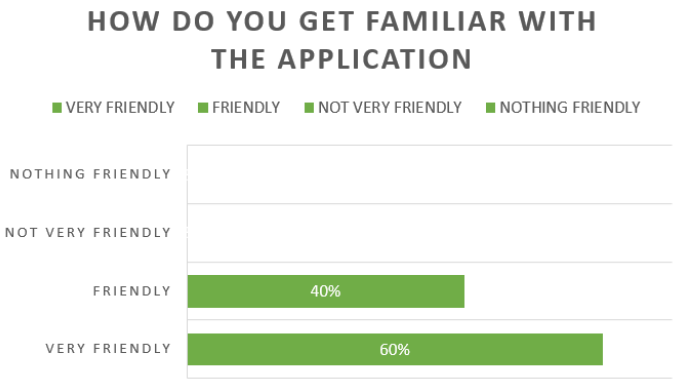


Fig. 8. Show the Results of Question 3.

- 60% define the application as very friendly and use.
- 40% defines the application friendly and use.
- The 0% defines how it is not very friendly and use of the application.
- 0% defines as not friendly and use of the application

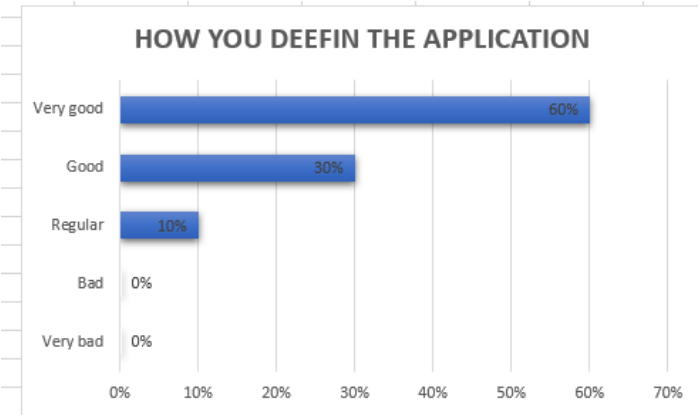


Fig. 9. Show the Results of Question 4.

Fig. 9 shows the results of question number 4 in which the respondent was asked, how do you define the application? Having 5 alternatives in an ascending way, they could define it as very good, good, fair, bad and very bad.The results obtained reflect the effort of the work team to achieve the final objective and the results were the following:

- 60% answered defined it as very good
- 30% defines the application as good.
- The 10% defines the application as regular.
- 0% defines it as bad.
- 0% defines it as very bad.

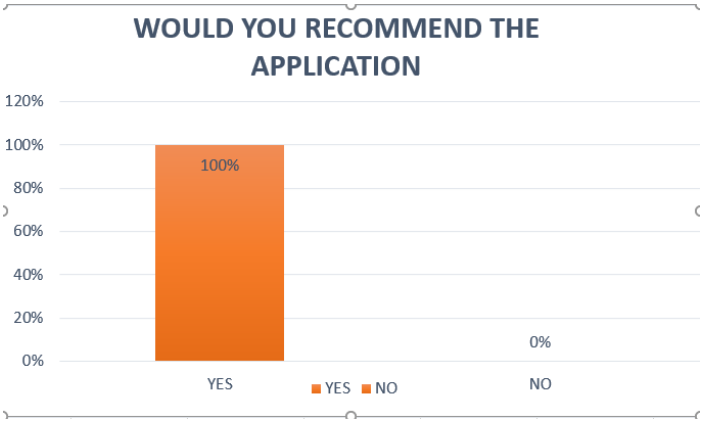


Fig. 10. Show the Results of Question 5.

In Fig. 10 shows the results of question number 5 in which the respondent was asked yes Would you recommend the application? The results were the following:

- 100% answered that they would recommend the application, which are very positive results for the project.
- 0% answered negatively.

B. About the Methodology

There are many alternatives to carry out research work which allows us to carry out large-scale projects. Scrum is a great and very significant alternative since it allows modifying in any part of the iterations without giving us complications, with reference to the [8] that use the Android operating system, surely the project in the future can be implemented to other mobile operating systems since it should not be limited to it.

As a work team we support the author [10], as it is an innovation system for society, in addition, it must be treated with the legal norms established by the executive as well as local and regional governments, who are responsible for its jurisdiction at the territorial level.

C. Prototype

For the creation of the prototype, the tool called Photoshop has been used, since being a graphic tool it allows to resemble the closest thing to reality. These prototypes were made according to the requirements of the system to be used as shown in the following figures.

However, there are tools such as: PropApp, Marvel, Proto.io, Fluid, Balsamiq that also serve to make prototypes for mobile applications that could be implemented in the future [21], however Photoshop was chosen since it was an easily mastered tool , but the possibilities of the other tools are not ruled out since they also provide facilities for the developer who has greater command or in which he feels more comfortable working [22].

Fig. 11. Shows the Design of the Prototype of User Story No. 2.

Fig. 11 shows the design of the prototype of User Story No. 2, where the user when entering the application will show him this interface. Within this interface, the user will log in and thus will be able to authenticate their entry with their username and password, taking them to the main interface. If you are entering for the first time, you have the option of registering (CKEKIN IN), clicking this button takes you to another interface of the application.

Fig. 12 shows the design of the prototype of User History No. 1 where the new user registers for the first and only time, he will have to enter his personal data such as name, surname, email, password, address, cell phone. This record is validated with a message to your entered email or to the entered cell phone.

Fig. 13 shows the design of the prototype of User History No. 3 where the application allows through geolocation to show the user his exact location where he is, displaying it on the map.

Fig. 14 shows the design of the prototype of User History No.5, in this interface the application allows to search for an address entered by the user, through geolocation it will show the predetermined route so that the user can go to the place where they want move.

Fig. 15 shows the design of the prototype of User Story No.12 where the user has a menu of options in a drop-down located in the upper left part of the application such as configuration, create routes, see their trips, log out.

Fig. 12. Shows the Design of the Prototype of User Story No. 1.

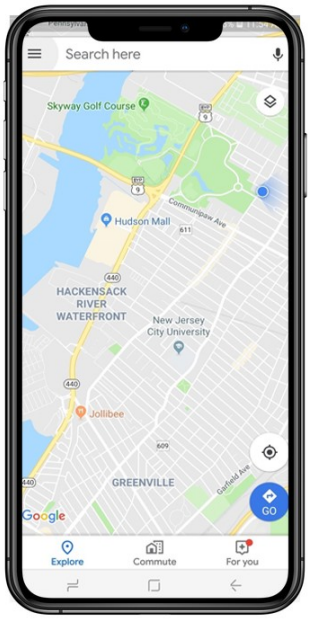


Fig. 13. Shows the Design of the Prototype of User Story No. 3.

Fig. 16 shows the design of the prototype of User History No.13 where the application shows the user the option to create their own route, it also allows them to save the route so that at another time they can access it and also allows them to share it with others users.

V. Conclusion

It is concluded that regarding the development of the mobile application, it shows all the functional requirements of the system, in addition to complying with the iterations that were proposed by the developers in the initial meeting of the research framework. The Scrum methodology allowed

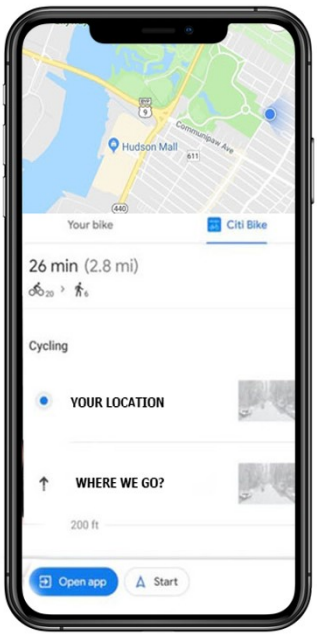


Fig. 14. Shows the Design of the Prototype of User Story No. 5.

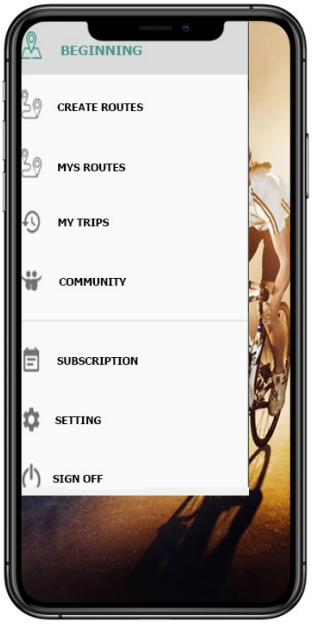


Fig. 15. Shows the Design of the Prototype of User Story No.12.

us to work with deliverables for each sprint, making quality prototypes for their development.

In its entirety in the survey carried out, people state that they would recommend the use of the application; that is to say that there is a satisfaction in the use they make of it. Some limitations were found, such as the little culture of the citizens for the use of bicycles on the routes; as well as the little support from the authorities to give priority to this issue.

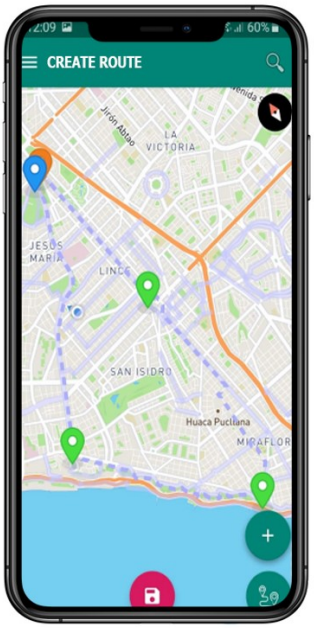


Fig. 16. Shows the Design of the Prototype of User Story No.13.

VI. Future Work

For future work, it is recommended to use this research work and implement new ideas such as the development of the application implementation for the different platforms and operating systems that exist in the Peruvian market, you can also implement the application to use it at the level national and not only in a sector limited to its use, this work will help a lot for the next investigations to be developed.The use of bicycles contributed to the citizens, since using bicycles allows to reduce pollution and stress, which would be the subject of another study in the future.

Acknowledgment

Recognition is made for the support provided by the University of Sciences and Humanities, through its Research Center, so that the research carried out can be achieved.

References

[1] M. Cabanillas-Carbonell, H. Paucarcaja-Ochoa, and O. Casazola-Cruz, "Analysis of the impact of ict for the management and control of public transport: A review of scientific literature from 2015 - 2020," in *2021 IST-Africa Conference (IST-Africa)*, 2021, pp. 1–9.

[2] A. C. Bazan Fachin, C. E. Calderon Mendoza, and C. S. Campos Salas, "Aplicativo móvil que ayudara a usuarios a trasladarse en bicicletas, scooters ya pie por rutas seguras y confiables: Move."

[3] J. A. Avalos Alfaro, J. J. Castro Palomino, V. L. Munoz Zumaran, D. A. Rubio Rotalde, and K. L. Silva Rivera, "Bicity-plataforma web y aplicativo móvil."

[4] S. R. Fajardo, L. F. Albán, and P. C. Guerrero, "Incidencia de las rutas de ciclismo en la demanda de turismo activo de naturaleza de los ciclistas que recorren la provincia del guayas. diseño de un aplicativo móvil de cicloturismo," *Revista Empresarial*, vol. 11, no. 44, pp. 7–12, 2017.

[5] A. Carrion-Silva, C. Diaz-Nunez, and L. Andrade-Arenas, "Admission exam web application prototype for blind people at the university of sciences and humanities," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111246

[6] A. Tupia-Astoray and L. Andrade-Arenas, "Implementation of an e-commerce system for the automation and improvement of commercial management at a business level," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120177

[7] M. Alvariño Torres and D. A. Leiva López, "Implementación de una aplicación móvil bajo la plataforma android para promover el uso de la bicicleta en lima metropolitana en el año 2019," 2020.

[8] R. Arias-Marreros, K. Nalvarte-Dionisio, and L. Andrade-Arenas, "Design of a mobile application for the learning of people with down syndrome through interactive games," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111187

[9] S. Olivi, L. Luciana, Y. Lozano, D. Martín, J. Luz, G. Cerrón, and O. Miryana, "Aplicación de movilidad urbana : Smartbikes aplicacón de movilidad urbana :," pp. 0–107, 2020.

[10] A. José, A. Rivera, V. Asesora, M. Marfil, F. Ballvé, M. Del, J. M. C. Torres, H. M. Cesar, and Z. Loayza, "El uso de la bicicleta como alternativa de transporte sostenible e inclusivo para Lima Metropolitana," p. 186, 2015.

[11] J. Alexander, M. Zamalloa, C. Descripción, and D. E. L. Problema, "Carrera Ingeniería Informática y de Sistemas Contenido," 2020.

[12] J. F. Cadavieco, "La interactividad de los dispositivos móviles geolocalizados , una nueva relación entre personas y cosas," no. October 2013, 2015.

[13] V. Gomero-Fanny, A. R. Bengy, and L. Andrade-Arenas, "Prototype of web system for organizations dedicated to e-commerce under the scrum methodology," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120152

[14] P. Deemer, G. Benefield, C. Larman, and B. Vodde, "Información básica de scrum," *California: Scrum Training Institute*, 2009.

[15] A. Ramos-Romero, B. Garcia-Yataco, and L. Andrade-Arenas, "Mobile application design with iot for environmental pollution awareness," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, pp. 566–572, 2021, cited By :3. [Online]. Available: https://doi.org/10.14569/IJACSA.2021.0120165

[16] J. Sandoval, *RESTful Java Web Services*. Packt Publishing, 2009, vol. 1.

[17] J. F. Cadavieco, "La interactividad de los dispositivos móviles geolocalizados , una nueva relación entre personas y cosas," no. October 2013, 2015.

[18] R. Elmasri and S. B. Navathe, *Fundamentals of Database System*, 2021.

[19] P. Bourhis, J. L. Reutter, and D. Vrgoč, "Json: Data model and query languages," *Information Systems*, vol. 89, p. 101478, 2020.

[20] I. Sudiartha, I. Indrayana, I. Suasnawa, S. Asri, and P. W. Sunu, "Data structure comparison between mysql relational database and firebase database nosql on mobile based tourist tracking application," in *Journal of Physics: Conference Series*, vol. 1569, no. 3. IOP Publishing, 2020, p. 032092.

[21] R. Leon-Ayala, G. G´omez-Cortez, and L. Andrade-Arenas, "Mobile application aimed at older adults to increase cognitive capacity," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0121297

[22] 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)*, 2020, pp. 1–4. [Online]. Available: https://doi.org/10.1109/EHB50910.2020.9280282

A Software Framework for Self-Organized Flocking System Motion Coordination Research

Fredy Martínez, Holman Montiel, Edwar Jacinto
Universidad Distrital
Francisco José de Caldas
Bogotá D.C., Colombia

***Abstract*—We describe and analyze the basic algorithms for the self-organization of a swarm of robots in coordinated motion as a flock of agents as a strategy for the solution of multi-agent tasks. This analysis allows us to postulate a simulation framework for such systems based on the behavioral rules that characterize the dynamics of these systems. The problem is approached from the perspective of autonomous navigation in an unknown but restricted and locally observable environment. The simulation framework allows defining individually the characteristics of the basic behaviors identified as fundamental to show a flocking behavior, as well as the specific characteristics of the navigation environment. It also allows the incorporation of different path planning approaches to enable the system to navigate the environment for different strategies, both geometric and reactive. The basic behaviors modeled include safe wandering, following, aggregation, dispersion, and homing, which interact to generate flocking behavior, i.e., the swarm aggregates, reach a stable formation and move in an organized fashion toward the target point. The framework concept follows the principle of constrained target tracking, which allows the problem to be solved similarly as a small robot with limited computation would solve it. It is shown that the algorithm and the framework that implements it are robust to the defined constraints and manage to generate the flocking behavior while accomplishing the navigation task. These results provide key guidelines for the implementation of these algorithms on real platforms.**

***Keywords*—Flocking; formation control; motion planning; multi-robot system; obstacle avoidance; swarm**

I. Introduction

A field of great interest in robotics poses the solution of problems not with a high-performance robot but with a group of robots, simpler in structure, but which can interact to behave as a single system [1], [2]. These systems are known as multi-agent, and seek to exploit the ability of biological systems such as flocks of birds [3], [4], schools of fish [5], [6], ants [7], [8] or aggregations of bacteria to solve complex tasks together [9], [10]. However, the control of these types of systems raises problems of high complexity given the characteristics of these dynamics. These are systems with self-organization capacity, which results as an emergent consequence of the system from basic behaviors of each of the agents [11]. The design of these basic behavioral rules that should generate the emergent behavior of the system is difficult. In addition, the robots, as a single system, must perform some task.

The flocking behavior presents interesting characteristics that make it of high interest for the design of artificial systems, particularly in problems of localization, search and rescue. This type of behavior has been observed in birds, is similar to schooling fish and swarming insects, and is characterized by a joint movement of the group without central coordination [12], [13]. The first basic rules of this dynamic were established in 1987 as alignment, cohesion, and separation [14]. Subsequently, in 2003, a mathematical model of the 1987 Reynolds rules was proposed taking advantage of the geometrical strategies and the theory of Artificial Potential Fields (APF) [15], [16]. In these works, it was demonstrated how potentials in the environment are not only able to guide the navigation of the robots to the target point, but also to form a homogeneous flock along with the navigation task [17]. It has been observed in different applications that much of the success of this dynamics lies in the local communication scheme between agents, which determines the ability to self-organize within the system [18], [19].

The dynamics of a flock of birds corresponds to that of randomized search algorithms [20]. These algorithms rely on a large number of agents distributed in the search space, which at the same time identify local information, maintain communication with their neighbors to jointly identify the most optimal solution option. In the case of the flock of agents, this solution corresponds to the region of the environment with the characteristics most similar to those defined in the search problem. These characteristics of the dynamics are what make it suitable for solving complex navigation problems, particularly in unknown and dynamic environments.

The complexity in the design of these systems lies in the fact that they can produce a system that is unable to solve the task either because the number of robots is too high or too low, if the local signal being tracked is too strong for the population size, or if the navigation environment is too complex [21]. These types of problems can drive the system to local minima, or impede the performance of the task. The motion planning of each robot within the system is linked to the structure of the system and the local information it can detect from the environment [22], [23]. Part of the information from the environment is transmitted to each robot from the movement of the system, which makes it robust to continuous changes of dynamics in the environment, but also dependent on the characteristics of the environment and the system for the success of the task. In this sense, the proper design of both the system and the behavioral policies of the agents is fundamental, particularly if we are looking for a system made up of agents with modest computational resources [24].

This research focuses on the development of multi-agent navigation strategies in this type of environment, guided by the identification of regions of interest in the environment [25].

These strategies must be following the task to be solved, so parameters such as system size, the distance between agents, the scope of the communication system, and characteristics of the basic behaviors must be different in each case [26]. Some assumptions are also made to simplify the model, but without moving it functionally away from the real prototypes. Among the initial assumptions, it is proposed to work with a single robotic platform (uniform system) with perfect displacement capabilities, constant values of the parameters throughout the development of the task, and discrete behavior of the agents in the sense that each action is triggered by a certain stimulus [27]. The objective is to achieve preliminary results that allow evaluating the success of the strategy and its possible implementation in real robots. Possible tasks in real applications include tracking pollution sources (static, dynamic, and multi-focal), intruder detection, or wildlife tracking.

One of the first developments in software tools to replicate the behavior of robot swarms was the one developed by Craig Reynolds to demonstrate the performance of his basic rules for flocking behavior in this type of multi-agent systems [14]. A later evolution by the same author was the OpenSteer project, a framework for implementing autonomous agent motions and behaviors [28]. These tools were originally intended for video game development, yet robotics has benefited from their use in demonstrating behavioral algorithms on them. The advantage (and disadvantage) in the use of this type of tool is the need for prior knowledge about the dynamics of multi-agent systems, a simulator for video games does not require such information, while a simulator dedicated to robotics requires it while allowing great versatility in the implementation of algorithms.

Perhaps the best-known robotic simulation tool is Player Project [29]. It is an open-source software specifically developed for robotics that allows a large number of controls and sensors to be incorporated into a client/server network interface. The results of these simulations can be visualized in the two-dimensional module Stage or the three-dimensional viewer Gazebo. Unfortunately, the last update of the project was made at the end of 2010. Still, Gazebo evolved independently, and today it is a dedicated high-performance simulation tool. Another project developed for robotics is Robot Virtual Worlds or RVW [30], which, although oriented more for robotics education purposes, under specific conditions can function as a simulation platform with its programming language. Other platforms worth mentioning include the CoppeliaSim project [31], and Webots from Cyberbotics Ltd [32], active projects, programmable in different languages, with a commercial focus.

We present the problem formulation in Section II. Here we give special attention to the basic behaviors that generate a flocking structure, and to the simplified representation assumed for the navigation problem. Section III presents the methodology followed for the construction of the framework, detailing not only the algorithm used but also the characteristics of its implementation. Section IV shows the results achieved in tests for controlled laboratory conditions, under which it is possible to perform experimental validation, and finally, the conclusions are presented in Section V.

II. Problem Statement

The flocking behaviors emerge in a multi-agent system as a consequence of a combination of basic or primitive behaviors executed independently by each of the agents. Among these primitive behaviors, five of them can be selected as the basis for the flocking structure [33]:

- Safe wandering: In the design of path planning solutions it is fundamental to guarantee the safe wandering of the robot along with the environment, this includes reducing collisions with other agents as well as with the obstacles and limits of the environment [34].
- Following: Robots must be able to establish their motion strategy from the motion of nearby robots. In the case of flock-based navigation strategies, each agent must identify neighboring agents, calculate its distance to them, and define its forward direction and speed to reduce interference in the system's motion [4].
- Aggregation: While agents must follow the movement of their neighbors, flocking behavior also requires that agents can dynamically assemble during navigation while maintaining a safe distance between them [35].
- Dispersion: Like aggregation, dispersion (self-localization of each robot in the system) turns out to be an important quality in autonomous coordination schemes, and is fundamental to the structure of the system throughout task development [36].
- Homing: This behavior allows each agent to move to the given target as part of the system task using sensed information in the environment [37].

That is, in a flock, each agent wants to stay close to the other agents (which it can detect), to do its best not to collide with them, and to move simultaneously towards the desired location. These behaviors make it possible to define a robust and well-organized flock. The objective of this research is to demonstrate that these basic behaviors allow generating a flocking behavior for an artificial system from fixed rules. This demonstration is done by simulation in a framework developed in Python for a system composed of TurtleBot 3 Burger robots from ROBOTIS. The goal is to scale the primitive behaviors to a large population of these agents.

Consequently, the problem is defined from the activation of n holonomic robots with known physical dimensions (not points) in an environment W unknown to the robots but partially observable from their sensors, and which is defined in a connected and compact two-dimensional plane $\left(W \subset \mathbb{R}^2\right)$. From this definition, it follows that all the constraints to which the TurtleBot 3 robotic platform can be subjected are integrable in positional constraints of the form:

$$f\left(q_1,\, q_2,\, q_3, \cdots, q_n; t\right) = 0 \tag{1}$$

where the variables q_i corresponds to the coordinates of the system.

Any typical environment W to be modeled by the framework contains in its interior a set of obstacles called O

that consists of regions inaccessible to the robot within W, where each of these regions is characterized by a closed and connected boundary. Therefore, the set O is also considered connected, finite, and piecewise analytic. An additional characteristic of each of the obstacles in O is that they are disjoint pairs of each other, so they do not share common points. The boundaries of W, denoted by ∂W, constrain the movement of the robots within the environment. In addition, the boundaries of the obstacles are also part of ∂W. The free space through which the robots can navigate is denoted by E and is defined as $W - O$.

Each of these robots (Fig. 1), according to its mechanical design, can be represented in the two-dimensional environment by a dish with a radius of 0.105 m (with center at the LiDAR sensor position) and an obstacle detection range (field of view of the LiDAR sensor) of 360 degrees with a range of 3.5 m. Other parameters derived from its design include a maximum forward speed of 0.22 m/s, and an acceptable range to define that it has reached a certain point in the environment of ± 0.5 m.

These robots have no explicit communication among themselves, only the ability to locate themselves and define their relative position concerning their neighbors (a basic type of local communication). The simulation framework assumes that the robots' sensing capability is perfect, that they are capable of perfect omnidirectional motion, and that they all follow the same navigation rules to define the path in W (Fig. 2).

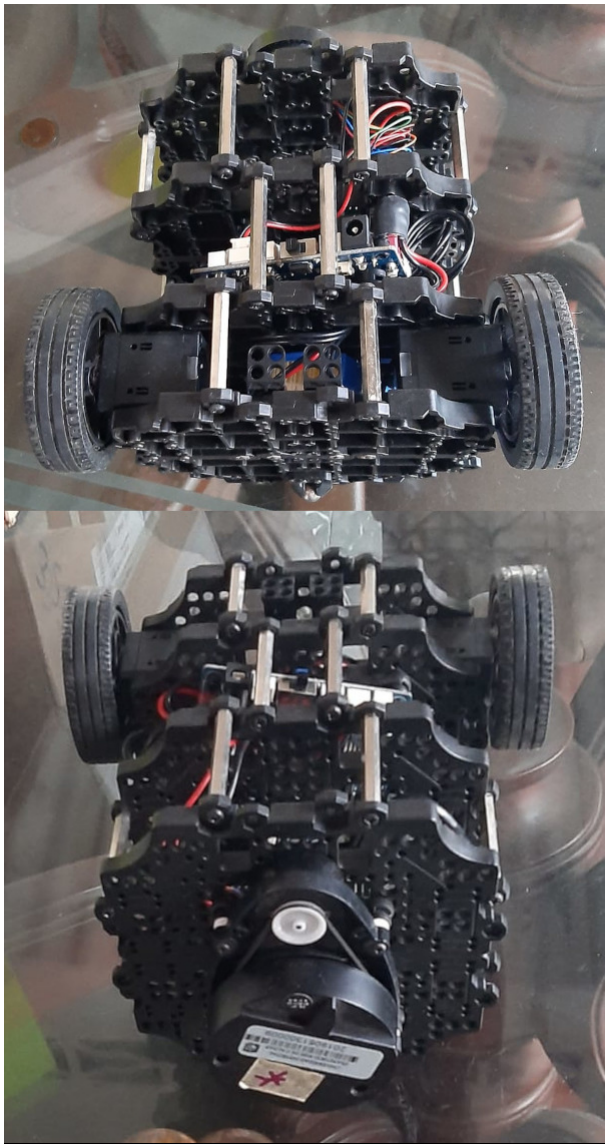


Fig. 1. TurtleBot 3 Burger Robot from ROBOTIS

III. Methods

The framework was developed in Python 3.7.12, with support for Numpy 1.19.5, Scipy 1.4.1, and Matplotlib 3.2.2. The tool allows simulating the movement of robots in the environment at a scaled relative speed, and the result is compressed into a video file. This is achieved with the Matplotlib animation library, included in the Matplotlib 1.1 version, which enables to obtain a visual demonstration of the behavior from the features programmed in the navigation algorithm. The base class of the animation tool is `matplotlib.animation.Animation`, on which the animation functionality is built. The interfaces of this tool are `TimedAnimation` and `FuncAnimation`, the latter is the one used in our framework. More details are provided below. From Numpy we use the linear algebra library `numpy.linalg.norm` to calculate the norms of the n-dimensional vectors. From Scipy we use the libraries `scipy.spatial.distance.pdist` to calculate the distances between points in the n-dimensional space, and `scipy.spatial.distance.squareform` that takes the previous results to form a square matrix of distances.

The first part of the code defines the global variables of the framework, which correspond mostly to user-configurable parameters according to the conditions of the problem to be simulated (Fig. 3 and Fig. 4). These variables include the population size of the system (how many robots will conform to the multi-agent system), the size of each of the robots (two-dimensional circular shape is assumed), the sensing range of the 360-degree distance sensors, the distance programmed in each robot to initiate the avoidance policy, the maximum speed, distance from the target to consider that the robot reached its destination, interval between simulation steps in simulation time, the initial position of the robot swarm, and dimensions of the simulation environment (the shape is always assumed to be two-dimensional rectangular, complex shapes in the environment can be achieved later with the definition of the obstacles O).

The second section of the code defines the navigation strategy to be followed by the robot swarm, i.e. how it will move in W to find the target area. In this part, we have facilitated the incorporation of several common algorithms, from the explicit definition of the route using navigation coordinates to the incorporation of geometric strategies such as Potential Field algorithm, Dijkstra, and A*, and even reactive strategies based on local sensing. Also in this section, the location of the target region is defined.

The third part of the initial configuration corresponds to the definition of the obstacles O. These are drawn on the environment using the `matplotlib.patches` library,

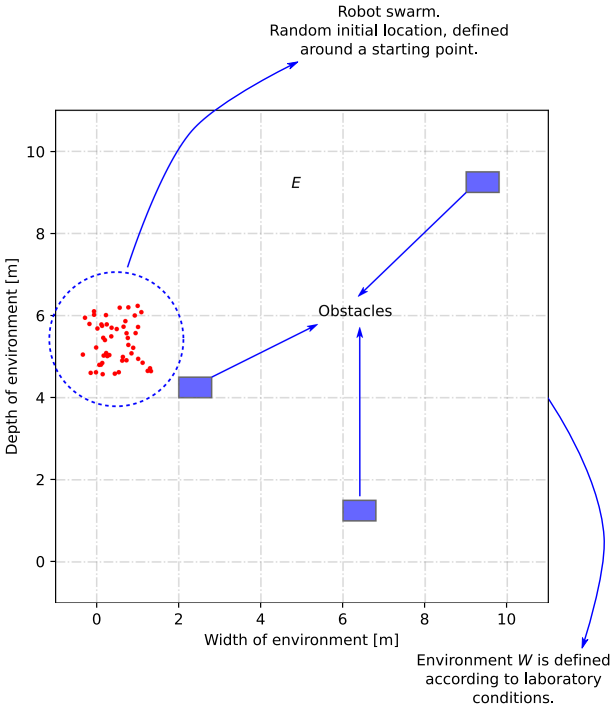


Fig. 2. Navigation Environment Defined in the Framework. The Definition of W, the Free Space E, the Obstacles, and the Swarm of Robots are Detailed

```
Initialize system
    - Number of agents
    - Agent size
    - Sensing range
    - Avoidance distance
    - Maximum speed
    - Final distance to target
    - Simulation window
    - Starting position
    - Environment dimensions
Initialize path planning
    - Navigation strategy
    - Location of target region
Obstacle configuration

BEGIN
    Build position and speed matrices
    For each agent i in population P do
        Background = init() → Obstacles
        Movement of agents = update()
            wandering()
            following()
            aggregation()
            dispersion()
            homing()
        Animate
    End For
End
```

Fig. 3. Pseudocode Detailing the Framework Structure

which corresponds to 2D objects defined by coordinates inside W. The coordinates are captured graphically with the help of a mouse pointer and converted to closed polygons that finally define E. It is also possible to define the coordinates of a rectangle, an option that is more versatile when looking for an analysis of sensitivity to the position of the obstacles. By default all obstacles in O are drawn in blue, to differentiate them from other elements of the simulation (agents are set in red and W boundaries in black).

With this information, we proceed to generate the initial position and velocity matrices of the system agents. In both cases, the values are generated randomly within the defined ranges. The initial state of the system is defined by scaling the environment and the robots for its visualization (the robots are represented as red-colored circles of proportional size to the environment). The figure is created with `matplotlib.pyplot.figure`, and all robots, obstacles, information labels, and a grid are added to facilitate position analysis. This configuration corresponds to the initial state of the system.

To perform the simulation the first thing to do is to initialize the robot speed arrangements. The speed of each robot can be kept constant at a percentage of the maximum value, or set to random in the same range. This is handled internally with a multiplier on the maximum speed between 0 and 1. This matrix is updated according to the motion policies applied to each robot. Animations in Matplotlib consist of three elements, an `init()` function that returns the background of each animation frame, an `update()` function that returns the figures that should appear in each background frame, and the code in charge of acquiring the animation object. In our case, the function `init()` loads the obstacles defined for the environment, and the function `update()` updates the position matrix from the previous matrix, the velocity matrix, and the result of the movement policies. The `animation.FuncAnimation` function takes as arguments the `init()` function, as well as the number of frames per iteration, the total animation interval, and some smoothing and updates instructions.

Each of the basic behaviors was implemented separately in functions that compute the velocity vector for each of the robots (each basis function for the entire system, from zero to n). The function for aggregation checks the readings from the distance sensors of each robot and adjusts the velocity vectors so that the robot moves towards its nearest neighbors. Distances to nearby robots are used to define the relative location of each robot in the environment, as well as its movement strategy [4]. A function is also implemented to establish attractors in the environment from the local readings, and the navigation strategy used. These attractors allow defining the velocity vector of each robot as if it were following a specific route. To avoid collisions, a function is defined that verifies the fulfillment of minimum distances between robots, obstacles, and environment limits, forcing the movement in a random direction when it detects a possible collision condition. These basic behaviors are combined to create more complex flocking behaviors. For example, the aggregation function and the attractor function combine to produce the following behavior and combined with the collision function they form a safe wandering behavior. Thus by combining homing, aggregation, and avoidance, the desired flocking behavior is achieved. In

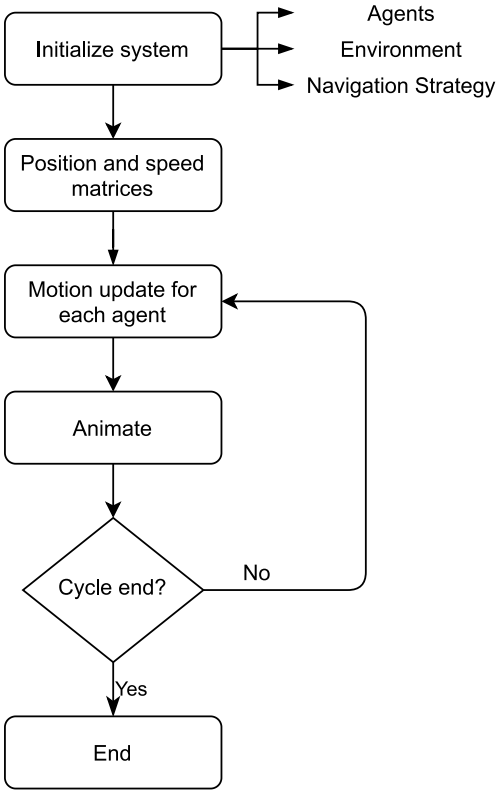


Fig. 4. Algorithm Flowchart

these combinations of basic behaviors, the effect of each is weighted to allow for a greater impact of one on the other, in most cases small homing versus high avoidance values and moderate aggregation produces the best flocking behavior.

IV. Results and Discussion

We have performed several tests with the framework for different conditions of the environment, the system, and the navigation strategy. The results shown below were performed for the TurtleBot 3 Burger robot, the platform on which we analyzed the flocking behavior. The characteristics of the environment (rectangular 10 m $\times$ 10 m, with three fixed obstacles) and the navigation strategy (reactive from intensity landmarks in the environment) were also kept constant. The varied parameters were population size and initial position of the system, with the intention not only to observe the self-organization in flocks but also to determine the performance of the system for a given navigation task concerning the population size. The system initialization parameters for these tests were as follows:

- Number of agents: between 5 and 100
- Agent size: circular with 0.105 m radius.
- Sensing range: 3.5 m
- Avoidance distance: 0.5 m
- Final distance to target: 0.5 m
- Simulation timestep: 0.01 s
- Starting position: Random in E
- Environment size: 10 m $\times$ 10 m

The navigation strategy forces the flock of robots to navigate the environment in a clockwise direction towards the lower right region of the environment. We evaluated the time it takes for the system to navigate to this region for different population sizes, from a flock of five agents to a system with 100 agents. Fig. 5 shows the capture of four such simulations, each with a different population size (20, 30, 40, and 50 agents). The starting point is randomly generated in E using a computer time-dependent variable seed. In the simulations, it is observed how the system self-organizes according to the basic behaviors, and after reaching equilibrium, it starts to move along the route without altering the formation, only in the event of encountering an obstacle, in which case the agent avoids it, and returns to the formation. A small animation of 20 s with these four cases can be seen in the following link:

https://youtu.be/R09mFqAb_-c

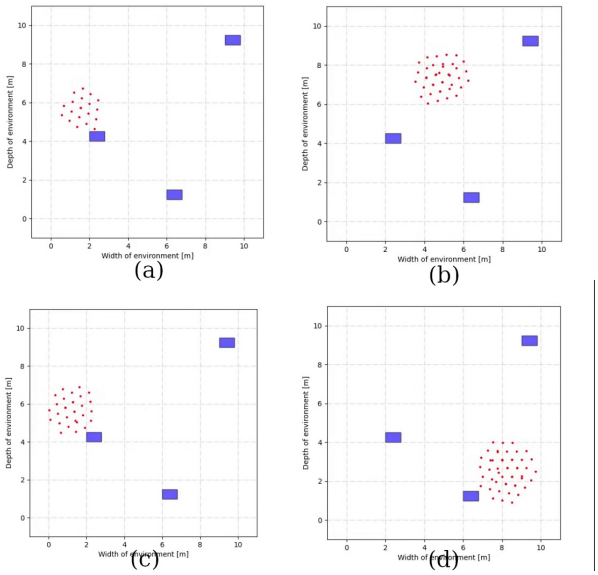


Fig. 5. Screenshots Corresponding to Four Simulations with Different Population Sizes Performing the Same Task. (a) 20 Agents, (b) 30 Agents, (c) 40 Agents, and (d) 50 Agents

The strength of the tool is observed when evaluating the performance of this type of system in the development of tasks, which is much more complex in implementations on real prototypes. To evaluate this, the above configuration was followed to assess the impact of population size on the total time required for task development. Since the navigation strategy relies on local readings, which may vary depending on the system agent detecting the landmarks, and according to the initial position of the system, statistical analysis with multiple simulations for multiple population sizes is needed to analyze the impact problem. Following the conditions of the previous simulations, the exercise was repeated for different population sizes: 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 agents. For each population size, 100 simulations were performed (1100 simulations in total) and the times in seconds that each case required were recorded. In the simulations, a 100% success rate

was obtained (in all cases the system reached the target region), but in some cases, the time required was an outlier (excessively long), which was also experimented with in real tests. The results are shown in Fig. 6, which shows a basic statistical analysis with median values by population size, quartiles, and excluded outliers.

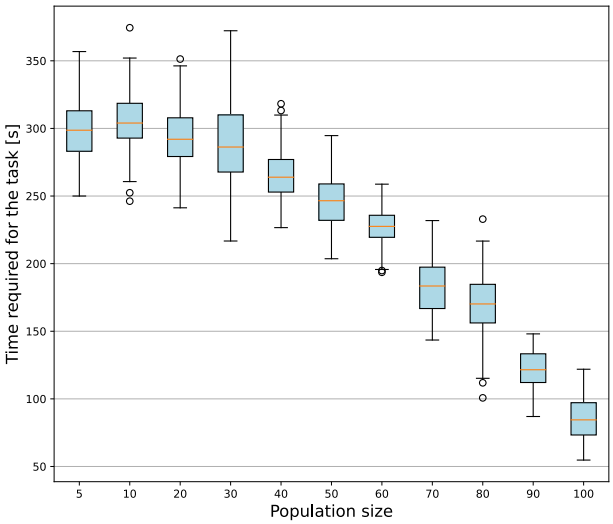


Fig. 6. Box and Whisker Plot of the Times for the Development of the Task for Different Population Sizes. Median, Quartiles, and Outliers are Detailed

The results show that beyond the fact that the system can perform the task, there is a relationship between the time required and the population size, for the particular working conditions (type of robot, size of the environment, and configuration of O) [25]. This relationship can be represented by a mathematical function, which could be useful for system sizing. This same exercise can be repeated with another set of obstacles, a different size environment, different speeds of the robots, or even some dynamic conditions making the obstacles change their position every so often. The framework allows to analysis and generalizes the behaviors of this kind of system, in a fast, economic and reliable way. These results can then be contrasted and scaled according to laboratory tests with some real robots.

V. Conclusion

In this paper, we propose a new simulation framework to replicate the flocking behavior of a swarm of robots, as a strategy for the efficient and safe evaluation of the performance of this type of system. The control scheme is designed based on the basic behaviors identified in the literature as fundamental for a flocking system: safe wandering, following, aggregation, dispersion, and homing. Each of these basic behaviors is replicated from each agent's sensor data, and weighted in conjunction with the navigation strategy to form the speed vector. We identify the weighted value of each basic behavior while allowing it to be adjusted according to the simulation needs. The framework also allows modeling different types of environments and robots, but the tests and calibrations were performed with the TurtleBot 3 Burger platform from Robotis. With this platform, the performance was tuned for a pair of agents, which was scaled to allow evaluating tens and hundreds of agents. In this sense, our framework allows us to adjust specific parameters such as robot size, sensing capacity, and maximum speed. The simulation performs an animation of the system using Python's Matplotlib library, which is exported to video. In this animation, the obstacles are placed in the background, while the agents are dynamically updated in the foreground. The code allows to implementation of a wide range of navigation strategies, both geometric from the global characteristics of the environment, as well as reactive navigation strategies based on local readings. The tool was verified for a simple navigation task, evaluating both the self-organizing capability of the system and the impact of system features on task performance. Future research on this tool includes the incorporation of other robotic platforms of the research group.

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] M. Talamali, A. Saha, J. Marshall, and A. Reina, "When less is more: Robot swarms adapt better to changes with constrained communication," *Science Robotics*, vol. 6, no. 56, pp. ST1–ST16, 2021.

[2] M. Otte, "An emergent group mind across a swarm of robots: Collective cognition and distributed sensing via a shared wireless neural network," *The International Journal of Robotics Research*, vol. 37, no. 9, pp. 1017–1061, 2018.

[3] H. Ling, G. Mclvor, K. Vaart, R. Vaughan, A. Thornton, and N. Ouellette, "Costs and benefits of social relationships in the collective motion of bird flocks," *Nature ecology & evolution*, vol. 3, no. 6, pp. 943–948, 2019.

[4] F. Martínez, "Minimalistic control scheme for the development of search tasks with flocks of robots," *Journal of Physics: Conference Series*, vol. 1993, no. 1, p. 012025, 2021.

[5] L. Demidova and A. Gorchakov, "Research and study of the hybrid algorithms based on the collective behavior of fish schools and classical optimization methods," *Algorithms*, vol. 13, no. 4, p. 85, 2020.

[6] L. Li, M. Nagy, J. Graving, J. Bak-Coleman, G. Xie, and I. Couzin, "Vortex phase matching as a strategy for schooling in robots and in fish," *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020.

[7] S. Jones, T. J. Czaczkes, A. J. Gallager, F. B. Oberhauser, E. Gourlay, and J. P. Bacon, "Copy when uncertain: lower light levels increase trail pheromone depositing and reliance on pheromone trails in ants," *Animal Behaviour*, vol. 156, no. 1, pp. 87–95, 2019.

[8] Q. Luo, H. Wang, Y. Zheng, and J. He, "Research on path planning of mobile robot based on improved ant colony algorithm," *Neural Computing and Applications*, vol. 32, no. 6, pp. 1555–1566, 2020.

[9] T. Trunk, , H. S. Khalil, and J. C. Leo, "Bacterial autoaggregation," *AIMS Microbiology*, vol. 4, no. 1, pp. 140–164, 2018.

[10] F. Martínez, "Robust eletronic hardware system based on quorum sensing," phdthesis, Universidad Nacional de Colombia, 2017.

[11] S. Zhang, M. Liu, X. Lei, Y. Huang, and F. Zhang, "Multi-target trapping with swarm robots based on pattern formation," *Robotics and Autonomous Systems*, vol. 106, no. 1, pp. 1–13, 2018.

[12] P. Zhu, W. Dai, W. Yao, J. Ma, Z. Zeng, and H. Lu, "Multi-robot flocking control based on deep reinforcement learning," *IEEE Access*, vol. 8, pp. 150 397–150 406, 2020.

[13] J. Cheng and B. Wang, "Flocking control of mobile robots with obstacle avoidance based on simulated annealing algorithm," *Mathematical Problems in Engineering*, vol. 2020, no. 1, pp. 1–9, 2020.

[14] C. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *Computer Graphics*, vol. 21, no. 4, pp. 25–34, 1987.

[15] H. Tanner, A. Jadbabaie, and G. Pappas, "Stable flocking of mobile agents. i. fixed topology," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*, 2003.

[16] ——, "Stable flocking of mobile agents. II. dynamic topology," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*, 2003.

[17] F. Martínez, *Robótica Autónoma: Arquitecturas Multiagente que Imitan Bacterias*. Universidad Distrital Francisco José de Caldas, 2021, vol. 1.

[18] C. Veitch, D. Render, and A. Aravind, "Ergodic flocking," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[19] E. Soria, F. Schiano, and D. Floreano, "The influence of limited visual sensing on the reynolds flocking algorithm," in *2019 Third IEEE International Conference on Robotic Computing (IRC)*, 2019.

[20] L. Li and A. Talwalkar, "Random search and reproducibility for neural architecture search," in *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, ser. 1, vol. 115, 2020, pp. 367–377.

[21] H. Wei and X. Chen, "Flocking for multiple subgroups of multi-agents with different social distancing," *IEEE Access*, vol. 8, pp. 164 705–164 716, 2020.

[22] O. Misir and L. Gökrem, "Flocking-based self-organized aggregation behavior method for swarm robotics," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 45, no. 4, pp. 1427–1444, 2021.

[23] L. Beaver and A. Malikopoulos, "An overview on optimal flocking," *Annual Reviews in Control*, vol. 51, no. 1, pp. 88–99, 2021.

[24] J. Benitez, L. Parra, and H. Montiel, "Diseño de plataformas robóticas diferenciales conectadas en topología mesh para tecnología zigbee en entornos cooperativos," *Tekhnê*, vol. 13, no. 2, pp. 13–18, 2016.

[25] A. Rendón, "Evaluation of autonomous navigation strategy based on reactive behavior for mobile robotic platforms," *Tekhnê*, vol. 12, no. 2, pp. 75–82, 2015.

[26] Y. Kambayashi, H. Yajima, T. Shyoji, R. Oikawa, and M. Takimoto, "Formation control of swarm robots using mobile agents," *Vietnam Journal of Computer Science*, vol. 06, no. 02, pp. 193–222, 2019.

[27] J. Castañeda and Y. Salguero, "Adjustment of visual identification algorithm for use in stand-alone robot navigation applications," *Tekhnê*, vol. 14, no. 1, pp. 73–86, 2017.

[28] U. Erra, B. Frola, and V. Scarano, "BehaveRT: A GPU-based library for autonomous characters," in *Motion in Games*, 2010, pp. 194–205.

[29] F. Martínez and D. Acero, "Autonomous navigation strategy for robot swarms using local communication," *Tecnura*, vol. 18, no. 39, p. 12, 2013.

[30] K. Takaya, T. Asai, V. Kroumov, and F. Smarandache, "Simulation environment for mobile robots testing using ROS and gazebo," in *2016 20th International Conference on System Theory, Control and Computing (ICSTCC)*, 2016, pp. 1–6.

[31] S. Rooban, S. Suraj, S. Vali, and N. Dhanush, "CoppeliaSim: Adaptable modular robot and its different locomotions simulation framework," *Materials Today: Proceedings*, vol. 2021, no. 1, pp. 1–6, 2021.

[32] L. Turkler, T. Akkan, and L. Akkan, "Control of swarm robotics in webots with PSO," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2021, pp. 1–6.

[33] M. Mataric, "Interaction and intelligent behavior," phdthesis, Massachusetts Institute of Technology, 1994. [Online]. Available: https://dspace.mit.edu/handle/1721.1/7343

[34] B. Tong, Z. Xiaoqing, L. Xiaoli, and R. Xiaogang, "A mobile robot path planning method based on safe pathfinding guidance," in *33rd Chinese Control and Decision Conference (CCDC 2021)*, 2021, pp. 1–6.

[35] O. Mısır, L. Gökrem, and S. M., "Fuzzy-based self organizing aggregation method for swarm robots," *Biosystems*, vol. 196, no. 1, p. 104187, 2020.

[36] A. Kshemkalyani, A. Rahaman, and G. Sharma, "Dispersion of mobile robots on grids," in *WALCOM: Algorithms and Computation*. Springer International Publishing, 2020, pp. 183–197.

[37] J. Xun, Z. Qidan, M. Junda, L. Peng, and Y. Tianhao, "Three landmark optimization strategies for mobile robot visual homing," *Sensors*, vol. 18, no. 10, p. 3180, 2018.

Trust-based Access Control Model with Quantification Method for Protecting Sensitive Attributes

Mohd Rafiz Salji[1], Nur Izura Udzir[2], Mohd Izuan Hafez Ninggal[3],
Nor Fazlida Mohd. Sani[4], Hamidah Ibrahim[5]
Faculty of Information Management,
Universiti Teknologi MARA, Malaysia[1]
Faculty of Computer Science and Information Technology,
Universiti Putra, Malaysia[2]

***Abstract*—The prevailing trend of the seamless digital collection has prompted privacy concerns to the organization. In enforcing the automation of privacy policies and laws, access control has been one of the most devoted subjects. Despite the recent advances in access control frameworks and models, there are still issues that hinder the implementation of successful access control. This paper illustrates the problem of the previous model which typically preserves data without explicitly considering the protection of sensitive attributes. This paper also highlights the drawback of the previous works which provides inaccurate calculation to specify user's trustworthiness. Therefore, a trust-based access control (TBAC) model is proposed to protect sensitive attributes. A quantification method that provides accurate calculation of the two user properties is also proposed, namely: seniority and behaviour to specify user's trustworthiness. Experiment have been conducted to compare the proposed quantification method and the previous quantification methods. The result shows that the proposed quantification method is stricter and accurate in specifying user's trustworthiness as compared to the previous works. Therefore, based on the result, this study resolves the issue of specifying the user's trustworthiness. This study also indicates that the issue of protecting sensitive attributes has been resolved.**

***Keywords*—Access control; trust-based access control; quantification method; sensitive attributes; privacy; privacy protection**

I. Introduction

Nowadays, information technology is growing exponentially, with an increasing number of hardware and software designed to make it easier for people to do their everyday work. This technology helps people to preserve their data privacy by using a wide variety of applications. Data can be collected, stored, and used for personal use or for work purposes. By using information technology, people can exchange data with the same interested party without any constraint of the boundary.

Data privacy is rapidly becoming one of the most crucial concerns in data management. People or customer is now more conscious about how their data are being protected by the organization. This awareness has been more highlighted when sharing and collecting data become seamless and prevalent by the omnipresence of Internet connection. In general, the organization collected, stored, and used customers' data for various purposes, and according to the Federal Trade Commission, U.S, 97 percent of websites were collected at least one type of identifying information such as name, e-mail address, or postal address of customers [1]. This could lead to misuse of customer data and less control over their privacy information. It may create privacy violations and fear to the customer [2]. Thus, data privacy should be protected in such a way that only authorized users can access the data. To protect data privacy, a relevant mechanism needs to be introduced by the company to build a solid trust with customers. The mechanism should be equipped with minimum requirements of reasonable access for privacy and security as stipulated in the Health Insurance Portability and Accountability Act (HIPAA, 1996).

There are several ways to protect data, but access control is the most common approach. It prohibits unauthorized access to the system resources as permitted by the policy [3], [4], [5], [6], [7], [8], [9], [10]. Trust-Based Access Control (TBAC) model is a popular access control paradigm that is influenced by an essential feature of human life that is trust. A user that is highly trusted would be given access to more resources as a part of this principle. However, trust is inconsistent in adapting to changing circumstances [11], [12]. Therefore, it is crucial to formulate an efficient access control model capable of capturing the complex essence of the scenario.

This paper addresses the issues of preserving sensitive attributes and determining the trustworthiness of the user. The previous TBAC models [13], [14], [15], which generally protect data without specifically focusing on protecting sensitive attributes, are outlined in this paper. Data is sensitive in nature, but sensitive attributes must be kept safe [16]. In general, data is categorized into three categories of attributes: de-identified, quasi identifier, and sensitive [17]. De-identified are the obvious identifying records that need to be hidden, such as the social security number. On the other hand, quasi identifier is a non-key attribute that has to be generalized before being published, such as race, age, and zip code, and finally, sensitive attributes are confidential data that belongs to a consumer privately, such as medical status and wages. According to the definitions of the three attributes, sensitive attributes require extremely restricted access in the system, with only trusted users permitted access to this attribute. In the previous models [13], [14], [15], trusted user, i.e., senior role, were granted more data access than the untrusted user, i.e.,

junior roles. However, these previous models did not mention which data could be accessed or not by trusted and untrusted users. This may lead the administrator to simply select any categories of data to be permitted or prohibited access by each trust level of the user. In this case, by not knowing which categories of data that is sensitive, the administrator may have the risk to disclose sensitive attributes to the untrusted user. Therefore, an access control model based on trust needs to be proposed to protect sensitive attributes.

Next, to access the resources, certain user properties need to be quantified to specify the user's trustworthiness whether the user is trusted or not to access it. Existing TBAC models [13], [14], [15] have been proposed to permit access to the resources of the system and introduce quantification methods by quantifying certain user properties to specify user's trustworthiness. If authorized user achieves highly trusted based on the calculation of user properties, they are permitted to access the data. However, these previous works provide an inaccurate assessment to specify a user's trustworthiness, which may cause the user who is still untrusted to become a trusted user. Therefore, an accurate quantification method needs to be proposed to calculate user properties to specify the user's trustworthiness. The measurement of the user properties with the detailed elements is also proposed to understand the process of calculation to specify the user's trustworthiness.

In summary, the main contributions of this paper are as follows:

1) An access control model based on trust is proposed to protect sensitive attributes.
2) A quantification method to calculate user properties to accurately specify user's trustworthiness is proposed.
3) A comprehensive set of calculations of user properties is proposed to understand the calculation process to specify the user's trustworthiness.

The rest of this paper is organized as follows: Section 2 provides the related works, while the user properties are discussed in Section 3. In Section 4, the proposed TBAC model framework is presented, while the calculation of user properties is described in Section 5. The proposed quantification method process is presented in Section 6, while the methodology is described in Section 7. The findings of this paper are explained in Section 8, and finally, Section 9 concludes the work.

II. Literature Review

In this section, the TBAC models which are closely related to the proposed model are discussed.

In the previous work, a trust-based RBAC model for pervasive computing systems has been proposed. Users' trustworthiness is evaluated by using the user properties, namely: experience and recommendation before they are assigned to roles or functions, i.e., senior role. The role is associated with trust. If the user achieves the minimum requirement of trust level set by an organization, the user can be assigned to that specific role and permitted to access the resources. A class of TBAC models with a very formal semantic that is expressed in a graph theory has been developed [13]. However, this previous model does not provide in detail how to quantify the user properties to determine the user's trustworthiness.

Previous work also proposes a privacy protection model to integrate trust management into access control [14]. The trust value of each requester is evaluated based on histories and recommendations. This model also includes purposes, obligations, and conditions that meet the requirements of modern cooperation, regulations, and privacy laws. However, this approach also does not include a thorough measurement of histories and recommendations to specify the user's trustworthiness to access the data.

The issue highlighted in the previous study [15] is the unreliability of the delegatee can cause disclosure of the delegator's privacy. Therefore, a multi-level delegation model with trust management has been proposed where delegation trustworthiness is organized in three levels: low, medium, and high. The more trustworthy the delegatee is, the more sensitive the delegation task able to be accessed by the delegatee. High denotes the person that has a higher level of trust. The low level of trust denotes the person that is less trusted, and finally, the medium level is the intermediate state. In this study, the evaluation of trust is based on the two interpretations of trust. First, the trust is based on the individual history and behaviour, called reliability trust, while the next interpretation is to capture trust by predicting trust trends in the forthcoming future, called future trust. However, this approach offers a general estimate of histories and recommendations to specify the user's trustworthiness.

A novel trust-based access control model in the cloud environment has been proposed to authorize the user and select the most trusted resources for the user [18]. The user trust value is evaluated based on the user behaviour parameter, and the resource trust value is evaluated based on the Service Level Agreement (SLA) parameter or the quality of service provided to the users. This model is compared with the existing Quality of Service (QOS) model and shows that the model performs better than the QoS model. However, the user trust value applied in the previous work is different as compared to the proposed work.

TrustRBAC is proposed based on trust and traditional role-based access model for single and multi-domain cloud environments [19]. The model calculates the direct trust and recommendation trust with security policies for both domains. The result shows that the TrustRBAC model effectively protects cloud users and secures its platform, thus achieving both the security and efficiency of the trust model. However, TrustRBAC model calculates using different properties as compared to the proposed model due to both works proposed in different environments.

A TBAC model is proposed with a comprehensive policy to specify the user's trustworthiness to access sensitive attributes and two properties are used to specify it, namely: seniority and behaviour [20]. Based on the calculation of both properties, if an authorized user achieves a higher level of trust (senior-with-trust), they can access sensitive attributes, otherwise, they are permitted to access data without sensitive attributes. However, this paper does not provide any test and validate the quantification method to specify the user's trustworthiness.

Finally, an access control model based on trust is pro-

posed for accessing data via cloud [21]. The level of user trustworthiness is classified into three levels: full, partial, and no view. The user who is trusted and semi-trusted is permitted to access a full and partial view of data, while the user who is untrusted is denied accessing data or no view. However, this paper does not provided any information on the calculation of user's trustworthiness.

All these works propose different approaches to protect the privacy of individuals by measuring different properties to specify the user's trustworthiness. The objective of this study is to preserve the sensitive attributes by using an access control model based on trust, and a quantification method is applied that provides an accurate measurement of the user properties to specify the user's trustworthiness. With this aim, this paper extends the previous works [13], [14], [15] by introducing an access control model based on trust that explicitly protects sensitive attributes, and in order to protect it, the user is calculated by using the quantification method to accurately specify the user's trustworthiness.

III. User Properties

In the TBAC model, certain user properties are required to determine the user's trust to access the resources. In the previous works [13], [14], [15], quantification methods have been introduced by calculating certain user properties to specify the user's trustworthiness to access the resources. However, the previous quantification methods have the limitation that provides an inaccurate formula to specify a user's trustworthiness that may cause the unauthorized user to become a trusted user to access the resources. In this study, due to the limitation of the previous works, a quantification method is proposed which provides the accurate calculation of the user properties, namely, seniority and behaviour to specify the user's trustworthiness. The discussion on seniority and behaviour is in the following sections.

A. Seniority

Seniority refers to the level of rank or position earned by a user or staff, which higher rank owns more priority compared to low. Based on previous works [13], [14], [15], experience or history is used to specify seniority which refers to the set of events or number of user activities that had occurred in the past within a certain period in which the user or trustee was involved and that the trustee has a recollection about. Examples of user activities include seminars, workshops, courses, and publications. However, this study is not only referring to the activities involved by the user, but the evidence that is relevant to calculate the seniority is also considered, for example, years in service, as this evidence is stated under the staffing policy [22]. Therefore, the evidence which is referred to the activities and relevant evidence is used to specify the seniority in this study. Seniority can be set in the role status attribute at the user's personal details. Two levels of user seniority are involved, junior (less trust) and senior (highly trust).

B. Behaviour

Behaviour refers to the user attitude or characteristic shown during their substantive service. Recommendation or trusted third-parties who have the knowledge about the user performance in service can be assigned by the administrator to evaluate the user behaviour [13]. In this study, the recommendation is set in the role trust attribute at the user's personal details. Three levels of user behaviour are involved in the proposed quantification method, mistrust (junior), trust (senior), and uncertainty (senior performing negative behaviour).

C. User's Trustworthiness and Access to the Resources

This section discusses the influence of user's trustworthiness in accessing the data especially sensitive attributes in the proposed model. If a user's seniority is assigned as a junior, the proposed work will automatically assign behaviour as mistrust. This is due to a junior referring to new staff, and mistrust refers to the staff that cannot be trusted. In this case, a user is still not achieving the minimum requirement of the seniority and behaviour set in the proposed quantification method. It denotes that a user is untrusted and is only permitted to access data without sensitive attributes. Next, a user also can be assigned the seniority as a senior, and behaviour is uncertainty. Previously, a user is assigned as senior with trust, but due to the user performing negative activities, for example, committing fraud, ignorance of obligation, and dishonest behaviour, unfortunately, an administrator has the right to change manually a user behaviour from trust to uncertainty. In this situation, a user has achieved a minimum requirement set in the proposed quantification method to become a senior, but the behaviour is set as uncertainty, which refers to a punishment for the user who performs wrongdoing. Therefore, a user is not permitted to access the sensitive attributes. Administrators, in this case, are the people at the top management level that are highly trusted to protect users' and customers' privacy. Finally, if user seniority is senior and behaviour is trust, a user is considered as a trusted user, and permitted to access the sensitive attributes. The influence of seniority and behaviour levels to authorize access to sensitive attributes is as shown in Fig. 1.

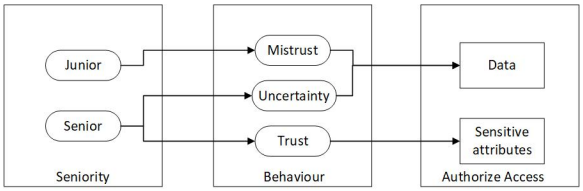


Fig. 1. User's Trustworthiness and Access to Sensitive Attributes

In this section, a proposed trust-based access control model is discussed. In general, to develop an access control model, three concepts are needed, access control model, policy, and mechanism [23], [10]. These three concepts are discussed in the proposed TBAC model framework as shown in Fig. 2.

The three main modules are discussed below.

1) Module 1: Access Control Policy
An access control policy is one of the concepts that need to be considered to implement access control. This policy is designed by organizations that are normally specific to their own use and may not be appropriate for other organizations [24]. In the

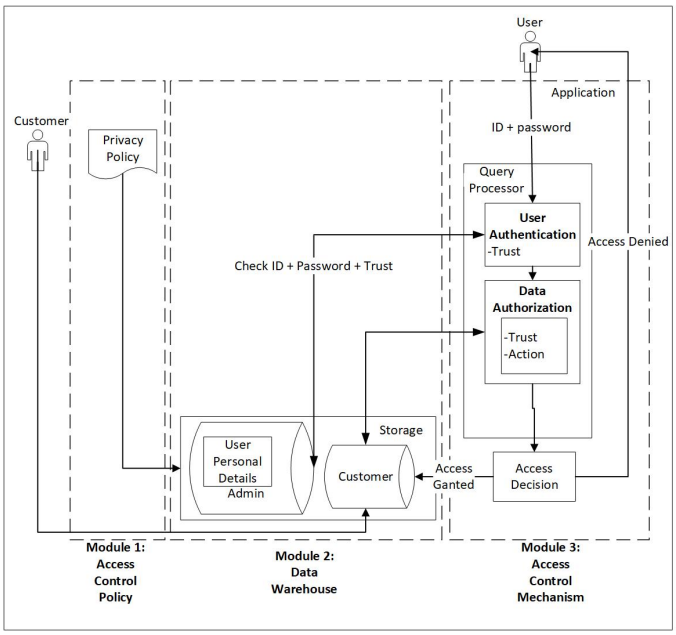


Fig. 2. The Proposed Trust-Based Access Control Model Framework

proposed model, access control policy issues are not part of the investigation.

2) Module 2: Data Warehouse
In this module, there are two databases involved in managing the proposed model, namely, the admin and customer database. These two databases are explained as follows:
   a) Admin database
   Admin database is used to collect the user data. In this database, the user personal details table is applied to store the user data. There are many particulars that can be collected from the user, but the user properties applied in the proposed model are the role status (user seniority) and role trust (user behaviour) attributes. These properties need to be identified before accessing sensitive attributes.
   b) Customer database
   Customer database store, maintain, and collect data from the customer. This database relates to the proposed model to permit authorized and trusted users only to access the customer data, particularly sensitive attributes.

3) Module 3: Access Control Mechanism
In access control, the data are protected by using two levels of access control mechanism, namely, user authentication and data authorization [25], [26], [27]. Normally, in the authentication stage, the user is authenticated based on username and password [28], [29]. However, due to privacy concerns by many parties, i.e., organization and customer, the expansion in terms of validating certain of the user properties must be considered to guarantee the correct user accesses the right data. As a result, the proposed model employs trust to authenticate the user to permit access to sensitive attributes.
In the next stage, the data are filtered based on certain user properties. If the user is trusted, sensitive attributes are rewarded to them, otherwise, the user is permitted to access data without sensitive attributes.

CALCULATION OF USER PROPERTIES

In order to specify the user's trustworthiness, a quantification method is required to quantify the user properties. In this section, the proposed quantification method is discussed to calculate the seniority and behaviour to specify the level of user's trustworthiness. The process of quantifying the seniority and behaviour in the proposed quantification method is described in the following sections.

D. Quantifying Seniority

In the proposed quantification method, the evidence is introduced to specify the user seniority. This evidence is stored in the user role history (URH) database and calculated by using the concept of weighing evidence [28] as shown in Fig. 3. The previous work [28] has suggested using weighing evidence to calculate the evidence that gives effect to the user's trustworthiness, however, the previous work has limitations to describe what is the user property used to be specified by the evidence. In this study, the weighing evidence is applied to calculate the evidence to specify the user seniority either senior or junior. The score of the evidence is stored in the URH database. The quantification of the evidence by using weighing evidence is discussed in the next section.

1) Weighing Evidence: In this study, weighing evidence is a method or decision process to quantify the evidence to specify the user seniority whether the user is senior or junior.

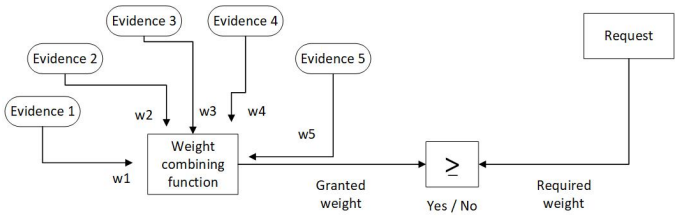


Fig. 3. Weighing Evidence

To quantify the evidence, the administrator needs to identify how many categories of evidence are to be set to calculate the seniority. However, the organization can determine the evidence as the different organization performs different evidence. For example, five categories of evidence are used to calculate the user Alice's seniority, i.e., years in service, seminars, workshops, courses, and publications. The value of calculating each category of evidence is between [0, 1] and the sum of this calculation is 1 [13]. In order to quantify the five categories of evidence, an administrator needs to decide the calculation on how to obtain the scores in each category. For example, the score of years in service can be based on how long a user works in an organization, e.g., a user obtains 0.1 mark for one year in service, and if a user has ten or more years in service, it means that the user obtains 1 mark or a full mark for years in the service category. The minimum required weight needs to be set by the administrator to specify whether the user is qualified to be a senior or not.

Let s denote the evidence and s needs to be calculated. The total sum of s is calculated as $(s_1 + \ldots + s_n)$. Then, the total sum of s is divided by a total number of evidence to obtain the result of user seniority us. The result is in the range of $[0, 1]$.

Hence, the administrator a must decide the minimum required weight of us. If the result of us is more than the required weight set by a, user u is assigned as a senior.

Assume the minimum required weight set by the administrator is 0.4. The calculation of user Alice's seniority is as follows: 1) Years in service = 0.5, 2) Seminars = 0.4, 3) Workshops = 0.6, 4) Courses = 0.3, 5) Publications = 0.7

The result of user Alice

```
   If seniority ≥ 0.4
   Result = senior
else
   Result = junior

(0.5 + 0.4 + 0.6 + 0.3 + 0.7) / 5 = 0.5
Result = senior
```

Based on the result, the user Alice's overall score is 0.5. This means that she is qualified as a senior.

E. Quantifying Behaviour

In this study, ten user behaviour categories are proposed to specify the user behaviour either trust or mistrust. Nine behaviour categories are proposed by Bruhn [30], and one category, self-discipline is proposed in this work as illustrated in Fig. 4. Self-discipline is included in this study as it is one of the user behaviour which is not included in the previous work [30], and it can be regarded as conscientiousness which implies a desire to do a task well and to take obligations to others seriously [31]. The justification for employing nine user behaviour categories in the proposed work is because they are based on a concept from prior work [30], and these categories were not undertaken in the computer domain. These categories are also used in the proposed work because the dataset collected and utilized in this model also uses the same categories. Subsequently, the reason for using self-discipline in the proposed work is because this category has been included in the dataset acquired from the Head of Studies Centers (HSCs). Therefore, these ten user behaviour categories are applied and quantified in the proposed quantification method to determine the user behaviour. The score of the ten user behaviour categories is stored in the recommendation database.

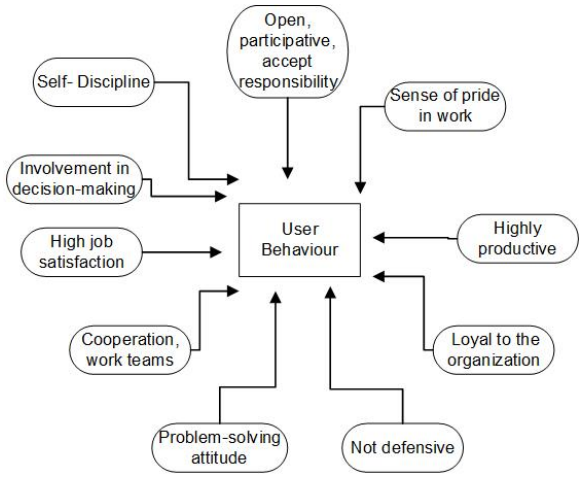


Fig. 4. User Behaviour Categories

The value of each category is between [0, 1] and the sum of these categories is 1 [13]. As mentioned earlier, the recommendation is applied in the previous works [13], [14], [15] to evaluate user behaviour. In this study, the recommendation is also used to evaluate the ten user behaviour categories. For example, if the mark of self-discipline is 0.1, the user obtains the lowest score, and if the score is 1 means the user obtains the highest score in that category. The minimum required weight needs to be set by the administrator to specify whether the user is qualified as trusted or mistrusted.

Let b denote the behaviour category and ten b needs to be quantified. The total sum of b is $(b_1 + \ldots + b_{10})$. Then, the total sum of b is divided by ten to obtain the result of a user behaviour ub. The result is in the range of $[0, 1]$. The ub is calculated as in Equation 1.

$$ub = \frac{1}{10} \sum_{i=1}^{10} b_i \tag{1}$$

Hence, the administrator a must decide the minimum required weight of ub. If the result of ub is more than the required weight set by a, user u is assigned as trust.

For example, assume the minimum required weight set by the administrator is 0.4. The user Alice's scores of behaviour is calculated as follows: 1) Open, participative, accept responsibility = 0.5, 2) Highly productive = 0.5, 3) Loyal to the organization = 0.5, 4) Not defensive = 0.5, 5) Cooperation, work teams = 0.5, 6) High job satisfaction = 0.5, 7) Problem-solving attitude = 0.5, 8) Involvement in decision-making = 0.5, 9) Sense of pride in work = 0.5, 10) Self-discipline = 0.5 The result of user Alice

```
   if behaviour ≥ 0.4
   Result = trust
else
   Result = mistrust

(0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 +
0.5 + 0.5 + 0.5) / 10 = 0.5
Result = trust
```

As a result, Alice's overall score is 0.5 and entitled as trust. In the previous work [32], the level of trust is introduced to identify the level of user's trustworthiness. However, each of the values is not set with the trust range, but the trust range is introduced in the notion of Vidyalakshmi et al. (2013). Therefore, in this proposed quantification method, the level of trust is introduced based on a combination of the previous works [32], [33] to identify the user level of trust, and this level of trust is as shown in Table I. This table indicates that the overall score of user Alice's behaviour is in Level 3, which is average.

TABLE I. LEVEL OF TRUST

Value	Meaning	Explanation	Trust Range
Level 0	Distrust Completely	Untrustworthy	0
Level 1	Ignorance	Cannot decide	0.1-0.19
Level 2	Minimal	Lowest trust	0.2-0.39
Level 3	Average	Mean trustworthiness	0.4-0.59
Level 4	Good	Trusted by major population	0.6-0.79
Level 5	Fully trust	Fully trustworthy	0.8-1

As mentioned earlier, three levels of user behaviour are proposed in this study, trust, mistrust, and uncertainty. Based on the quantification of user behaviour, only two levels of user behaviour are involved, trust and mistrust. As explained earlier, uncertainty is changed manually from the trust by an administrator. Previously, the user is set as a senior-with-trust, and as the user is performing negative activities, the user is set as a senior-with-uncertainty. Therefore, the user is not allowed to access sensitive attributes.

F. Computation of Trustworthiness

In this study, the user's seniority and behaviour are quantified to determine the level of user's trustworthiness. If the calculation of seniority attains the minimum required weight set by the administrator, but the calculation of behaviour does not achieve minimum requirement or vice versa, the user is not assigned as senior-with-trust. In this case, both properties should achieve the minimum requirement set by the administrator to become a trusted user, or else the user is still set as a junior-with-mistrust. For example, based on the previous sections (refer Section III - D1 & E), the user Alice's trustworthiness needs to be specified. The result of the user that is trusted is set as 1, while the untrusted user is stated as 0. The calculation to specify Alice's trustworthiness is as follows:

The result of user Alice

```
   if (Seniority ≥ 0.4) & (Behaviour ≥
0.4)
   Result = 1
else
   Result = 0

(Seniority = 0.5) & (Behaviour = 0.5)
Result = 1
```

As a result, both properties of user Alice has achieved the minimum requirement and she has qualified to become a trusted user. Based on the result, Alice is allowed to access sensitive attributes in the proposed model. In the next section, the function and process in the proposed quantification method to specify the user's trustworthiness are discussed.

IV. PROPOSED QUANTIFICATION METHOD

In this section, the framework of the proposed quantification method is designed to discuss each function and process. The framework is composed of two main modules, and it is shown in Fig. 5. The two main modules are as follows:

1) Data Warehouse
2) Request and Calculation

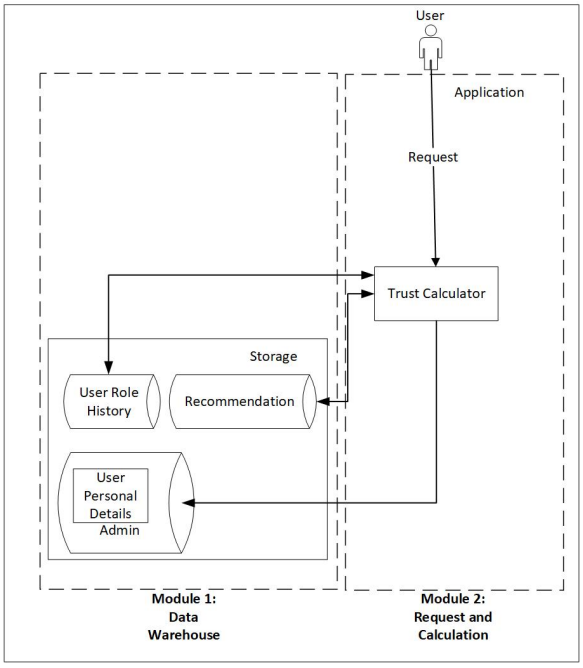


Fig. 5. The Proposed Quantification Method Framework

Now, the two main modules are discussed.

1) Data Warehouse
This module refers to the three databases, user role history (URH), recommendation, and admin database. These databases are used to collect different types of

data. The explanation of these three databases is as follows:

a) User role history (URH) database
In the previous works [14], [15], [13], the URH database is applied to store the score of activities. In the proposed quantification method, the score of evidence is stored in the same database to quantify the evidence to specify the user seniority whether the user is senior or junior. If a user requested to be a trusted user, the score of a user's evidence is calculated in the trust calculator to specify the user seniority. After specifying the user seniority, the result is stored in the admin database.

b) Recommendation database
In the previous works [14], [13], a recommendation database is introduced to store the score of user's behaviour. This score is supplied by the recommendation or evaluator who knows the user. In the proposed quantification method, the recommendation database is also used to store the score of the user behaviour based on the ten user behaviour categories to specify whether the user is trusted or mistrusted. If a user requests to become a trusted user, the score of ten user behaviour categories is calculated in the trust calculator to specify the user behaviour. The result is stored in the admin database after specifying the user behaviour.

c) Admin database
After the user seniority and behaviour are quantified, the result is stated in this database at the user personal details table. The user personal details table includes the user information and some of the attributes are assigned for user authentication. In the proposed quantification method, role status and role trust attributes are used to state the result based on the quantification of the user seniority and behaviour. The role status attribute is used to state the user's seniority whether the user is junior or senior, while the role trust attribute is used to state the result of user behaviour either trust, mistrust, or uncertainty. Based on both attributes, the results can be used in the proposed model to identify user's trustworthiness to access sensitive attributes.

2) Request and Calculation
In this module, the request to calculate user's trustworthiness and calculation of the user properties is discussed as follows:

a) User: Refers to new staff or a user who requested to become a trusted user. A new user is set as a junior-with-mistrust or untrusted user and requested to be set as a senior-with-trust or trusted user. A user needs to request from the system to quantify user properties to specify the user's trustworthiness. The quantification procedure will then be carried out automatically. A senior-with-uncertainty user can also be requested to be set as a senior-with-trust where the proposed work will re-evaluate user behaviour by re-calculating the ten user behaviour categories to specify whether the user is trusted or maintained to uncertainty. However, if the user performs wrongdoing, the administrator is responsible for manually changing user trust from senior-with-trust to senior-with-uncertainty. The proposed quantification method obtains the score of user evidence at the URH database to calculate the user seniority, while the score of ten user behaviour categories is taken from the recommendation database to calculate the user behaviour. Based on the calculation of both properties, the result is stored in the admin database.

b) Trust calculator: In this process, the user seniority and behaviour are quantified to specify the user's trustworthiness. If the result of user seniority is senior, but the user behaviour is mistrust, or the user seniority is junior, but the user behaviour is trust, the user is still considered as an untrusted user. The user is considered trusted if the user is set as a senior-with-trust.

V. Methodology

The proposed quantification method needs to be developed to test whether it can be used to specify or not a user's trustworthiness by quantifying the user properties, seniority, and behaviour. Both properties are quantified by using a dataset. Next, the proposed quantification method needs to be validated by comparing the calculation of the user properties to specify the user's trustworthiness with the previous quantification methods [13], [14], [15]. The proposed quantification method shows the calculation of the user seniority and behaviour, while the previous quantification methods present the quantification of the user history | experience and recommendation. The dataset is explained in the following section.

A. Dataset

To test the proposed quantification method, the dataset is required to calculate the user seniority and behaviour to specify the user's trustworthiness. This study uses a dataset that contains staff data that refer to their performance assessment, and the data need to be prepared every year-end. This dataset is applied in this study due to containing the data of user seniority and behaviour to evaluate the lecturer's performance. It is acquired from the Head of Studies Centers (HSCs) at Universiti Teknologi MARA, Sarawak who are responsible to assess the performance of the lecturer under the HSC's responsibility.

In this study, 48 original user data (refer to Appendix) are collected from the HSCs. The data are then expanded as synthetic data to show the variety of results from the different numbers of user data. The function of random between (RANDBETWEEN) is applied to increase the amount of data

until 500 users. Then, this study shows the pattern of the proposed quantification method results between the original and synthetic data. Next, the proposed quantification method is compared with the previous works [13], [14], [15] to highlight the differences of results between the proposed work and the previous works.

B. Proposed Quantification Method

This section discusses the proposed quantification method in quantifying the user seniority and behaviour to specify the user's trustworthiness. To show the result of user's trustworthiness, four tests are conducted where the first test used 48 original data, while, the following three tests [7], [34] are conducted by using synthetic data with a different number of users, i.e, 100, 300, and 500 to understand the pattern of the result between the original and synthetic data in the proposed and previous works. The calculation results of both user properties to quantify user's trustworthiness are discussed in the following sections.

1) Test on Quantification of Seniority: In this study, to test the quantification of user seniority, the result is in the range of [0, 1]. Since the total score of activities in the dataset is set as 10 marks, therefore it needs to be calculated as follows: (score of user's seniority / 10). The minimum requirement to be a senior is set at 0.8 marks based on the pattern that the majority of the users attain that minimum score. To calculate the user seniority, the score of **User 3** as shown in Table II is used as an example. It shows that User 3 has reached the minimum requirement and qualified as a senior.

TABLE II. LEVEL OF USERS' SENIORITY

No.	Name	Activity / Contribution	Total	Rolestatus
1	User 1	8	0.8	senior
2	User 2	9	0.9	senior
3	User 3	8	0.8	senior
4	User 4	6	0.6	junior
5	User 5	7	0.7	junior
6	User 6	8	0.8	senior
7	User 7	8	0.8	senior
8	User 8	8	0.8	senior
9	User 9	8	0.8	senior
10	User 10	9	0.9	senior

As mentioned earlier, four tests are conducted to obtain the results of the quantification of users' seniority in the proposed work. These results of the four tests are applied to specify the user's trustworthiness in the following section. The results are as shown in Table III.

2) Test on Quantification of Behaviour: To test the quantification of user behaviour, the result is set in the range of [0, 1]. The total score of each category is set as 10 marks. This study has ten categories of user behaviour, therefore, the total score of all categories is set as 100 marks. The calculation of the user behaviour is set as follows: (sum of all categories / 100). The minimum requirement to attain a level of trust is set at 0.8 marks due to the majority of users attaining that minimum score. The score of **User 3** as shown in Table IV is applied as an example to calculate the user behaviour. It shows that User 3 has accomplished the minimum requirement where the score is 0.9. Based on Table I, User 3 is in level 5 means the user is fully trusted.

TABLE III. RESULT OF THE USERS' SENIORITY

Test	Senior	Junior	Total
1	36	12	48
2	54	46	100
3	138	162	300
4	229	271	500

Four tests are conducted to quantify the user behaviour in the proposed work. These tests are utilized to specify the user's trustworthiness in the next section. The results of the four tests are shown in Table V.

3) Result of User's Trustworthiness: In the previous sections (refer Section V - B1 & B2), four tests are conducted to calculate the users' seniority and behaviour and the results of both properties have been specified. Then, this section enlightens how to specify the users' trustworthiness. The result of the user that is trusted is set as 1, while the untrusted user is stated as 0. The result of users' trustworthiness is shown in Table VI. For example, in Table VI, the score of **User 3** are as follows:

The result of User 3

```
if (Seniority ≥ 0.8) & (Behaviour ≥ 0.8)
   Result = 1
else
   Result = 0

(Seniority = 0.8) & (Behaviour = 0.9)
Result = 1
```

As a result, both properties of User 3 have achieved the minimum requirement and are qualified to become trusted user. Based on the result, User 3 is allowed to access sensitive attributes in the proposed model.

Based on the four tests to specify the users' seniority and behaviour, the results are utilized to specify the user's trustworthiness in the proposed quantification method. The result of the four tests to specify the users' trustworthiness are as shown in Table VII.

Next, the discussion is on the calculation of the previous quantification methods [13], [14], [15], and later validate the proposed quantification method by comparing the result of the calculation with the previous quantification methods.

C. Existing Quantification Methods

The calculation of the previous quantification methods is as shown in Table VIII. By using the same score of both properties as in the proposed quantification method, then the combination of both properties in the previous quantification methods is calculated as follows: (Experience + Recommendation) / 2. The result of the user that is trusted is set as 1, while the untrusted user is stated as 0. For example, in Table VIII, the score of **User 3** are as follows:

The result of User 3

```
   if (Average [Experience | History +
Recommendation] 0.8)
   Result = 1
```

TABLE IV. QUANTIFICATION AND LEVEL OF USERS' BEHAVIOUR

No	Name	open	productive	loyalty	not defensive	cooperation	job satisfaction	problem solver	decision maker	sense of pride	discipline	Total	Activity	Total
1	User 1	9	9	9	9	9	9	9	9	9	9	0.9	8	0.8
2	User 2	9	9	9	9	9	9	9	9	9	9	0.9	9	0.9
3	User 3	9	9	9	9	9	9	9	8	9	9	0.9	8	0.8
4	User 4	8	8	8	8	9	8	9	8	8	9	0.8	6	0.6
5	User 5	8	8	9	9	8	9	9	8	9	9	0.9	7	0.7
6	User 6	9	8	9	9	9	9	9	8	9	9	0.9	8	0.8
7	User 7	9	8	9	9	8	9	9	8	9	9	0.9	8	0.8
8	User 8	9	8	9	9	9	9	9	8	9	9	0.9	8	0.8
9	User 9	8	8	9	9	8	8	9	8	9	9	0.9	8	0.8
10	User 10	9	9	9	9	9	9	9	8	9	9	0.9	9	0.9

TABLE V. RESULT OF THE USERS' BEHAVIOUR

Test	Trust	Mistrust	Total
1	46	2	48
2	79	21	100
3	207	93	300
4	343	157	500

TABLE VI. RESULT OF USERS' TRUSTWORTHINESS

No	Name	**TOTAL**	rolestatus	**TOTAL**	roletrust	**RESULT**
1	User 1	0.8	senior	0.9	trust	1
2	User 2	0.9	senior	0.9	trust	1
3	User 3	0.8	senior	0.9	trust	1
4	User 4	0.6	junior	0.8	trust	0
5	User 5	0.7	junior	0.9	trust	0
6	User 6	0.8	senior	0.9	trust	1
7	User 7	0.8	senior	0.9	trust	1
8	User 8	0.8	senior	0.9	trust	1
9	User 9	0.8	senior	0.9	trust	1
10	User 10	0.9	senior	0.9	trust	1

```
else
   Result = 0

(0.8 + 0.9) / 2 = 0.9
Result = 1
```

Based on the result of User 3 in the previous quantification methods, by merging the score of both properties, the user has achieved the minimum requirement to become a trusted user.

Similar to the proposed quantification methods, the four tests are conducted to determine user's trustworthiness in the previous quantification methods. These results are used to compare with the proposed quantification method in the next section. The results of the four tests are shown in Table IX.

VI. FINDINGS

In this section, the proposed quantification method needs to be validated by comparing the calculation with the previous methods. Table X shows the comparison of the result between the proposed and previous quantification methods [13], [14], [15]. Based on the results of the previous examples (refer Section V - B3 & C), User 3 in the proposed quantification method and previous quantification methods achieve the same result which is 1 or qualified as a trusted user. In this case, there is no difference between both methods because both of them achieve the same results, even the calculation of both methods is different. However, the result of User 5 is different as compared to User 3. The discussion on **User 5** are as follows:

TABLE VII. RESULT OF THE USERS' TRUSTWORTHINESS IN THE PROPOSED QUANTIFICATION METHOD

Test	Trusted User	Untrusted User	Total
1	36	12	48
2	45	55	100
3	101	199	300
4	166	334	500

TABLE VIII. RESULT OF USERS' TRUSTWORTHINESS DATA FOR THE PREVIOUS QUANTIFICATION METHODS

No	Name	Experience / history / seniority	Recommendation / behaviour	TOTAL	RESULT
1	User 1	0.8	0.9	0.9	1
2	User 2	0.9	0.9	0.9	1
3	User 3	0.8	0.9	0.9	1
4	User 4	0.6	0.8	0.7	0
5	User 5	0.7	0.9	0.8	1
6	User 6	0.8	0.9	0.8	1
7	User 7	0.8	0.9	0.8	1
8	User 8	0.8	0.9	0.8	1
9	User 9	0.8	0.9	0.8	1
10	User 10	0.9	0.9	0.9	1

TABLE IX. RESULT OF THE USERS' TRUSTWORTHINESS IN THE EXISTING QUANTIFICATION METHODS

Test	Trusted User	Untrusted User	Total
1	40	8	48
2	66	44	100
3	171	129	300
4	298	202	500

Result in the **Previous Quantification Methods**:

```
   if (Average [Experience | History +
Recommendation] ≥ 0.8)
   Result = 1
else
   Result = 0

(0.7 + 0.9) / 2 = 0.8
Result = 1
```

Result in the **Proposed Quantification Method**:

```
   if (Seniority ≥ 0.8) & (Behaviour ≥
0.8)
   Result = 1
else
```

```
   Result = 0

(Seniority = 0.7) & (Behaviour = 0.9)
Result = 0
```

Based on the results of User 5, both methods show different results where the previous quantification methods achieve the minimum requirement to become a trusted user, while the proposed work shows the result is 0 or untrusted user. In the previous quantification methods, by combining the amount of both properties, User 5 achieves the minimum requirement to become a trusted user even the user seniority is still a junior. However, in the proposed quantification method, the rule is both user properties must achieve the minimum requirement, but unfortunately, User 5 in the proposed work does not achieve the minimum requirement of seniority to become a senior. Therefore, the result of User 5 is the untrusted user. These comparisons show the calculation in the proposed quantification method is stricter and accurate as compared to the previous quantification methods to be a trusted user.

In order to validate the proposed work, the results of the four tests of specifying the users' trustworthiness in the proposed work and previous works need to be compared. Based on Tables VII and IX, the results in Test 1 show the previous works and proposed work have many trusted users as compared to untrusted users. However, previous works have more trusted users, which are 40 people or 83.3% than the proposed work is only 36 people or 75%.

In Test 2, the result of a trusted user in the previous works and proposed work is different as compared to Test 1. The previous works show 66 people or 66% are trusted users, while the proposed work shows untrusted users are more than trusted users where trusted users recorded as only 45 people out of 100 or 45%. Although the percentage of the trusted users in the previous works declined as compared to Test 1, the previous works maintain records as a higher number of trusted users than the proposed work.

Test 3 also shows the same trend where the number of trusted users in the previous works and proposed work which are 171 people or 57% and 101 people or 33.7% decreasing as compared to Test 1 and Test 2. However, the previous works show the same pattern with more trusted users as compared to the proposed work.

Finally, in Test 4, the previous quantification methods show 298 users or 59.6% with an increase in the number of trusted users as compared to Test 3, while the proposed quantification method shows the same trend of decreasing number of trusted users that is only 166 users or 33.2%. However, Test 4 also shows the same trend with the proposed quantification method obtaining a smaller number of trusted users as compared to previous works.

To compare the pattern of the trusted user in Test 1 (original data) and three tests by using synthetic data, the results show the same pattern which the number of trusted users in the original and synthetic data is smaller in the proposed work as compared to the previous works. The reason for a smaller number of trusted users in the proposed work is because the proposed work provides strict rules which to ensure both properties achieve a minimum requirement to be a trusted user. The rule in the proposed work is accurate as compared to previous works because the proposed work ensures only the user who achieves the minimum requirement of both properties is qualified as a trusted user. As compared to the previous works, both properties are combined without considering both properties have achieved the minimum requirements or not. If achieves the minimum requirement, the user is entitled as a trusted user. Therefore, this comparison shows that the proposed quantification method is stricter and accurate in specifying user's trustworthiness as compared to the previous works.

The proposed method's strict rule achieves the notion of privacy by restricting data access to only authorized users [35], [36]. Even while previous work restricted access to authorized users in this scenario, the technique is insufficient to properly identify trusted users as compared to the proposed method, which allowed only trusted users to access sensitive attributes. Therefore, a strict rule is required in the proposed method to produce an accurate result.

In this study, there are three differences between the proposed quantification method and previous quantification methods [13], [14], [15].

1) Previous works applied the quantification methods to quantify certain properties to specify the user's trustworthiness to access data. While the proposed work uses a comprehensive set of quantification method which use the user seniority and behaviour to specify the user's trustworthiness, and later in the proposed access control model, the user's trustworthiness can only be identified to access sensitive attributes.
2) Previous works have a limitation which provides inaccurate calculation to quantify the user properties. However, this study proposes the quantification method with stricter and accurate calculations to specify the user's trustworthiness.
3) The score of experience and recommendation [13] and history and recommendation [14], [15] have been calculated and merged to specify the user's trustworthiness. However, the proposed work does not combine the score of seniority and behaviour to specify the user's trustworthiness. Therefore, the proposed work provides better calculation because the proposed work has to make sure both properties achieve a minimum requirement to become a trusted user.

VII. Conclusion

In this paper, a TBAC model has been proposed to protect sensitive attributes. A quantification method also has been proposed by providing accurate measurement of the two user properties, namely: seniority and behaviour to specify the user's trustworthiness. A detailed measurement of user properties is also proposed to understand the process of specifying the user's trustworthiness. Test and validation of the proposed quantification method have been conducted to prove that it can be used to specify the user's trustworthiness and compare it with the previous quantification methods. The result shows that the proposed quantification method is stricter and accurate in specifying user's trustworthiness as compared to the previous

TABLE X. RESULT OF THE PROPOSED AND PREVIOUS QUANTIFICATION METHODS

No	Name	Experience / History / Seniority	Rolestatus	Recommendation / Behaviour	Roletrust	RESULT - Proposed Method	TOTAL - Previous Methods	RESULT - Previous Methods
1	User 1	0.8	senior	0.9	trust	1	0.9	1
2	User 2	0.9	senior	0.9	trust	1	0.9	1
3	User 3	0.8	senior	0.9	trust	1	0.9	1
4	User 4	0.6	junior	0.8	trust	0	0.7	0
5	User 5	0.7	junior	0.9	trust	0	0.8	1
6	User 6	0.8	senior	0.9	trust	1	0.8	1
7	User 7	0.8	senior	0.9	trust	1	0.8	1
8	User 8	0.8	senior	0.9	trust	1	0.8	1
9	User 9	0.8	senior	0.9	trust	1	0.8	1
10	User 10	0.9	senior	0.9	trust	1	0.9	1

works. Based on the result, the issue of the previous works [13], [14], [15] have limitation which provides inaccurate calculation to specify user's trustworthiness has been solved. The issue of the previous access control models based on trust which focuses on generally protecting data without considering specifically protecting sensitive attributes also has been solved.

In future work, further development needs to be considered. First, many different types of access control models are employed to preserve privacy, such as blockchain-based access control [37], cloud-based access control [38], provenance-based access control model [39], and situation-based access control [40]. Therefore, this is an opportunity for researcher to develop alternative access control models to address the challenge of keeping privacy, particularly sensitive attributes. Next, the suggested quantification method may be adapted to specify authorized users or subjects to access resources in another environment, such as blockchain or cloud.

REFERENCES

[1] ANSI, "American national standard for information technology role based access control," *ANSI INCITS*, pp. 359–2004, February 2004.

[2] J.-W. Byun, E. Bertino, and N. Li, "Purpose based access control of complex data for privacy protection," in *Proceedings of the Tenth ACM Symposium on Access Control Models and Technologies*, ser. SACMAT '05. New York, NY, USA: ACM, 2005, pp. 102–110. [Online]. Available: http://doi.acm.org/10.1145/1063979.1063998

[3] C. Bertolissi and M. Fernández, "A metamodel of access control for distributed environments: Applications and properties," *Information and Computation*, 2014.

[4] J. Crampton and J. Sellwood, "Path conditions and principal matching: A new approach to access control," in *Proceedings of the 19th ACM Symposium on Access Control Models and Technologies*. ACM, 2014, pp. 187–198.

[5] R. Sandhu, D. Ferraiolo, and R. Kuhn, "The NIST model for role-based access control: Towards a unified standard," in *ACM Workshop on Role-based Access Control*, vol. 2000, 2000.

[6] P. C. Hung, "Towards a privacy access control model for e-healthcare services." in *Third Annual Conference on Privacy, Security and Trust, October 12-14, 2005, The Fairmont Algonquin, St. Andrews, New Brunswick, Canada, Proceedings*, 2005.

[7] A. Kayes, J. Han, and A. Colman, "A semantic policy framework for context-aware access control applications," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*, July 2013, pp. 753–762.

[8] A. Lazouski, F. Martinelli, and P. Mori, "Usage control in computer security: A survey," *Computer Science Review*, vol. 4, no. 2, pp. 81–99, 2010.

[9] S. Ruj, M. Stojmenovic, and A. Nayak, "Privacy preserving access control with authentication for securing data in clouds," in *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*. IEEE, 2012, pp. 556–563.

[10] P. Samarati, "Protecting respondents identities in microdata release," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 13, no. 6, pp. 1010–1027, 2001.

[11] S. Vahabli and R. Ravanmehr, "A novel trust-based access control for social networks using fuzzy systems," *World Wide Web*, vol. 22, no. 6, pp. 2241–2265, 2019.

[12] B. Zhao, C. Xiao, Y. Zhang, P. Zhai, and Z. Wang, "Assessment of recommendation trust for access control in open networks," *Cluster Computing*, vol. 22, no. 1, pp. 565–571, 2019.

[13] M. Toahchoodee, R. Abdunabi, I. Ray, and I. Ray, "A trust-based access control model for pervasive computing applications," in *Data and Applications Security XXIII*. Springer, 2009, pp. 307–314.

[14] M. Li, H. Wang, and D. Ross, "Trust-based access control for privacy protection in collaborative environment," in *e-Business Engineering, 2009. ICEBE'09. IEEE International Conference on*. IEEE, 2009, pp. 425–430.

[15] M. Li, X. Sun, H. Wang, and Y. Zhang, "Multi-level delegations with trust management in access control systems," *Journal of Intelligent Information Systems*, vol. 39, no. 3, pp. 611–626, 2012.

[16] N. Maheshwarkar, K. Pathak, and N. S. Choudhari, "K-anonymity model for multiple sensitive attributes," *International Journal of Computer Applications (IJCA)*, 2012.

[17] L. Sweeney, "K-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[18] P. K. Behera and P. M. Khilar, "A novel trust based access control model for cloud environment," in *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*. Springer, 2017, pp. 285–295.

[19] C. Uikey and D. Bhilare, "TrustRBAC: Trust role based access control model in multi-domain cloud environments," in *Information, Communication, Instrumentation and Control (ICICIC), 2017 International Conference on*. IEEE, 2017, pp. 1–7.

[20] M. R. Salji, N. I. Udzir, M. I. H. Ninggal, N. F. M. Sani, and H. Ibrahim, "Performance-aware trust-based access control for protecting sensitive attributes," in *International Conference on Soft Computing and Data Mining*. Springer, 2016, pp. 560–569.

[21] A. Singh, U. Chandra, S. Kumar, and K. Chatterjee, "A secure access control model for e-health cloud," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 2329–2334.

[22] S. Ganesan and B. A. Weitz, "The impact of staffing policies on retail buyer job attitudes and behaviors," *Journal of Retailing*, vol. 72, no. 1, pp. 31–56, 1996.

[23] S. Castano and E. Ferrari, "Protecting datasources over the web: Policies, models and mechanisms," in *Web-Powered Databases*. IGI Global, 2003, pp. 299–330.

[24] N. Abdul Ghani, "Credential purpose-based access control for personal data protection in web-based applications," Ph.D. dissertation, Universiti Teknologi Malaysia, Faculty of Computing, 2013.

[25] T. Ercan and M. Yıldız, "Semantic access control for corporate mobile devices," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2010, pp. 198–207.

[26] B. Lampson, M. Abadi, M. Burrows, and E. Wobber, "Authentication in distributed systems: Theory and practice," *ACM Transactions on Computer Systems (TOCS)*, vol. 10, no. 4, pp. 265–310, 1992.

[27] T. Singh and R. Kumar, "Database and information security concerns," *International Journal of Computer Science & Technology.*, vol. 4, no. 2, pp. 211–215, 2011.

[28] D. Gollmann, "From access control to trust management, and back–a petition," in *IFIP International Conference on Trust Management*. Springer, 2011, pp. 1–8.

[29] S. Harris, *Mike Meyers' CISSP Certification Passport*. McGraw-Hill/Osborne, 2002.

[30] J. G. Bruhn, *Trust and the Health of Organizations*. Springer Science & Business Media, 2001.

[31] S. Bai, B. Hao, A. Li, S. Yuan, R. Gao, and T. Zhu, "Predicting big five personality traits of microblog users," in *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*. IEEE Computer Society, 2013, pp. 501–508.

[32] M. Kim, J. Seo, S. Noh, and S. Han, "Identity management-based social trust model for mediating information sharing and privacy enhancement," *Security and Communication Networks*, vol. 5, no. 8, pp. 887–897, 2012.

[33] B. Vidyalakshmi, R. K. Wong, and C.-H. Chi, "Decentralized trust driven access control for mobile content sharing," in *Big Data (BigData Congress), 2013 IEEE International Congress on*. IEEE, 2013, pp. 239–246.

[34] A. Kayes, J. Han, and A. Colman, "An ontological framework for situation-aware access control of software services," *Information Systems*, 2015.

[35] C. Esposito, A. De Santis, G. Tortora, H. Chang, and K.-K. R. Choo, "Blockchain: A panacea for healthcare cloud-based data security and privacy?" *IEEE Cloud Computing*, vol. 5, no. 1, pp. 31–37, 2018.

[36] C. S. Powers, P. Ashley, and M. Schunter, "Privacy promises, access control, and privacy management. enforcing privacy throughout an enterprise by extending access control," in *Proceedings. Third International Symposium on Electronic Commerce,*. IEEE, 2002, pp. 13–21.

[37] C. Yang, L. Tan, N. Shi, B. Xu, Y. Cao, and K. Yu, "Authprivacychain: A blockchain-based access control framework with privacy protection in cloud," *IEEE Access*, vol. 8, pp. 70 604–70 615, 2020.

[38] A. A. Almtrf, "Cloud-based access control to preserve privacy in academic web services," Ph.D. dissertation, Oakland University, 2020.

[39] A. Bates, B. Mood, M. Valafar, and K. Butler, "Towards secure provenance-based access control in cloud environments," in *Proceedings of the third ACM conference on Data and application security and privacy*. ACM, 2013, pp. 277–284.

[40] D. Beimel and M. Peleg, "Using owl and swrl to represent and reason with situation-based access control policies," *Data & Knowledge Engineering*, vol. 70, no. 6, pp. 596–615, 2011.

APPENDIX

No	Name	open	productive	loyalty	not defensive	cooperation	job satisfaction	problem solver	decision maker	sense of pride	discipline	Total	Activity	Total
1	User 1	9	9	9	9	9	9	9	9	9	9	0.9	8	0.8
2	User 2	9	9	9	9	9	9	9	9	9	9	0.9	9	0.9
3	User 3	9	9	9	9	9	9	9	8	9	9	0.9	8	0.8
4	User 4	8	8	8	8	9	8	9	8	8	9	0.8	6	0.6
5	User 5	8	8	9	9	8	9	9	8	9	9	0.9	7	0.7
6	User 6	9	8	9	9	9	9	9	8	9	9	0.9	8	0.8
7	User 7	9	8	9	9	8	9	9	8	9	9	0.9	8	0.8
8	User 8	9	8	9	9	9	9	9	8	9	9	0.9	8	0.8
9	User 9	8	8	9	9	8	8	9	8	9	9	0.9	8	0.8
10	User 10	9	9	9	9	9	9	9	8	9	9	0.9	9	0.9
11	User 11	8	8	9	9	9	8	9	8	9	9	0.9	6	0.6
12	User 12	9	9	9	9	9	9	9	8	9	9	0.9	8	0.8
13	User 13	7	6	8	9	6	8	8	6	8	9	0.8	5	0.5
14	User 14	8	8	9	9	8	8	9	8	9	9	0.9	6	0.6
15	User 15	8	8	9	9	9	9	9	8	9	9	0.9	8	0.8
16	User 16	8	8	9	9	9	9	9	8	9	9	0.9	9	0.9
17	User 17	8	8	9	9	9	9	9	8	9	9	0.9	8	0.8
18	User 18	8	8	9	9	9	9	9	8	9	9	0.9	8	0.8
19	User 19	8	8	9	9	8	8	9	8	8	9	0.8	6	0.6
20	User 20	8	8	9	9	9	8	9	8	9	9	0.9	8	0.8
21	User 21	8	6	9	8	6	8	8	6	8	9	0.8	6	0.6
22	User 22	7	6	8	9	6	8	7	6	8	9	0.7	5	0.5
23	User 23	6	6	8	9	6	8	7	6	8	9	0.7	4	0.4
24	User 24	10	7	10	10	8	10	7	7	10	9	0.9	8	0.8
25	User 25	9	7	10	10	8	9	8	8	9	10	0.9	8	0.8
26	User 26	10	8	9	10	9	9	8	9	9	10	0.9	8	0.8
27	User 27	10	9	10	10	9	10	7	8	10	10	0.9	8	0.8
28	User 28	10	9	10	10	9	9	8	8	10	10	0.9	8	0.8
29	User 29	10	7	10	10	8	9	7	7	9	10	0.9	8	0.8
30	User 30	10	7	10	10	8	9	8	8	9	10	0.9	8	0.8
31	User 31	9	7	9	10	8	9	7	8	9	10	0.9	8	0.8
32	User 32	10	7	10	10	8	9	8	8	9	10	0.9	8	0.8
33	User 33	8	7	9	10	7	8	7	7	8	10	0.8	7	0.7
34	User 34	9	7	10	10	9	10	8	9	10	10	0.9	8	0.8
35	User 35	9	9	10	10	8	10	9	8	10	10	0.9	8	0.8
36	User 36	9	7	10	10	7	10	9	9	10	10	0.9	9	0.9
37	User 37	9	7	9	10	8	9	9	8	9	10	0.9	9	0.9
38	User 38	9	7	10	10	9	9	9	8	9	10	0.9	9	0.9
39	User 39	9	7	10	10	8	9	8	8	9	10	0.9	8	0.8
40	User 40	9	7	10	10	9	10	8	8	10	10	0.9	9	0.9
41	User 41	10	7	10	10	9	9	7	7	9	10	0.9	8	0.8
42	User 42	9	7	9	10	7	7	7	7	7	10	0.8	7	0.7
43	User 43	9	7	9	10	7	7	7	7	7	10	0.8	7	0.7
44	User 44	9	7	9	10	8	9	8	8	9	10	0.9	8	0.8
45	User 45	9	7	10	10	8	9	8	8	9	10	0.9	8	0.8
46	User 46	9	7	10	10	9	10	9	9	9	10	0.9	8	0.8
47	User 47	9	7	8	10	8	9	8	8	9	10	0.9	8	0.8
48	User 48	9	7	7	10	7	7	8	7	7	10	0.8	10	1.0

Feature based Entailment Recognition for Malayalam Language Texts

Sara Renjit
Department of Computer Science
Cochin University of Science and Technology
Kerala, India

Sumam Mary Idicula
Department of Computer Science
Muthoot Institute of Technology and Science
Kerala, India

***Abstract*—Textual entailment is a relationship between two text fragments, namely, text/premise and hypothesis. It has applications in question answering systems, multi-document summarization, information retrieval systems, and social network analysis. In the era of the digital world, recognizing semantic variability is important in understanding inferences in texts. The texts are either in the form of sentences, posts, tweets, or user experiences. Hence understanding inferences from customer experiences helps companies in customer segmentation. The availability of digital information is ever-growing with textual data in almost all languages, including low resource languages. This work deals with various machine learning approaches applied to textual entailment recognition or natural language inference for Malayalam, a South Indian low resource language. A performance-based analysis using machine learning classification techniques such as Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, AdaBoost, and Naive Bayes is done for the MaNLI (Malayalam Natural Language Inference) dataset. Different lexical and surface-level features are used for this binary and multiclass classification. With the increasing size of the dataset, there is a drop in the performance of feature-based classification. A comparison of feature-based models with deep learning approaches highlights this inference. The main focus here is the feature-based analysis with 14 different features and its comparison, essential to any NLP classification problem.**

Keywords*—Textual entailment; natural language inference; Malayalam language; machine learning; deep learning

I. Introduction

Textual entailment (TE), also called natural language inference (NLI) is a relationship between a pair of sentences. It identifies the similarity between the sentences based on their inferential semantic content. A text is said to entail another sentence, called a hypothesis, if the hypothesis has its semantic content derived from the text. In the same way, the text contradicts the hypothesis if the semantic content of the hypothetical sentence is just opposite to the text. Both sentences remain neutral to each other if the hypothesis derives zero information from the text.

A classical definition for entailment is that a text t entails hypothesis h if h is true in every circumstance of a possible world in which t is true. This definition is too strict in applying to real-world applications. An applied definition says that text t entails hypothesis h if human reading t infers that h is most likely true. Mathematically computable definition for text entailment is provided as hypothesis h is entailed by text t if P(h is true $|$t) $>$ P(h is true), where P(h is true $|$t) is the Entailment Confidence [1].

Semantic variability in expressions is an essential factor in any natural language processing application. NLI is also a necessary sub-task for almost all NLP applications such as multi-document summarization, question answering systems, information extraction, information retrieval. In multi-document summarization, the redundant sentences are identified using entailments, and those sentences can be removed. The answer to a question can be evaluated based on its entailment to the reference answer in the question answering system. In information extraction and retrieval systems, the text should entail the extracted information.

Natural language inference also finds application in analysis of user tweets, posts and experiences in social networks, where people share their thoughts, experiences in the form of texts in various languages. These texts are useful to relate between users by analysing inferences (entailment, contradiction and neutral) between the texts. This helps in customer segmentation, product analysis from the customer viewpoint as well as in recommender systems.

As information is available in digital text form in almost all languages, recognizing entailment is important for almost all languages. Text entailment is recognized in various languages, namely, English, French, Spanish, Italian, Japanese, Hindi, Swahili, Urdu. Very few works are reported for the Malayalam language.

In this work, we classify entailments for Malayalam, a South-Asian language from the Dravidian family. Malayalam is the language officially used and spoken in the state of Kerala. This language has its origin from the Dravidian scripts of Tamil. The language has various dialects, agglutinations, and inflectional word forms used in different parts of the state. This language also has very few resources in terms of datasets and other language processing applications and falls in the class of low resource languages.

The main contributions in this work includes:

1) The application of machine learning methods for Malayalam language textual entailment recognition, which is not attempted so far and also required for current literature and future research in this area.

2) A comparison between machine learning and deep learning approaches for Malayalam language entailment recognition.

3) The limitations of feature based methods with increasing dataset size.

4) An inference that deep learning without explicit feature-based engineering helped in more accurate classification for datasets of larger size.

The rest of the article is organized as follows: Section II describes the related literature in English and other languages. Challenges and contributions in Malayalam language for entailment is provided in Section III. Textual entailment for the Malayalam language using feature set is detailed in Section IV. The experimental evaluations are in Section V. Section VI discuss the results and Section VII concludes the work.

II. Related Work

Textual entailment has its inception in 2005 as PASCAL (Pattern Analysis, Statistical Modelling, and Computational Learning) challenge programme 'Recognizing Textual Entailment (RTE)' to develop systems that can recognize inferences from text fragments across various applications like multi-document summarization, information retrieval, information extraction and question answering systems.

In 2008, PASCAL RTE became a track at the Text Analysis Conference organized by NIST (National Institute of Standards and Technology), which brought different NLP communities to work on the textual entailment application scenarios. The earliest approaches for determining textual entailment include bag of words, logic-based reasoning, lexical entailment, machine learning methods, and graph matching [2].

The English language: The challenge started for the English language, and all major works are implemented in English language using RTE(Recognizing Textual Entailment), SNLI (Stanford Natural Language Inference) [3], MNLI (Multi-genre Natural Language Inference) [4] and XNLI (Cross-lingual Natural Language Inference) datasets. Lexical and syntactic similarity based entailment classification is done using rule-based similarity features such as unigram, skip-gram, longest common subsequence, stemming, subject-subject comparison, subject-verb, object-verb comparison [5].

RTE datasets were used to train and test these systems. Entailment recognition is also attempted by resolving anaphoras in sentence pairs [6]. [7] does similarity metrics-based recognition of entailments in the text, where features like cosine similarity, unigram match, Jaccard similarity, dice similarity, overlap, harmonic mean, and machine translation evaluation metrics, namely BLEU and METEOR, are used for machine learning. Following are the other approaches:

Bag of Words: In this approach, both text and hypothesis are represented as a collection of words. Every word from the hypothesis collection is compared with every word from the text collection. If the match between T and H is more than a preset threshold, then the sentence pair is classified as entailment, else, not entailment. It ignores the word order, syntax, and semantics of the sentences.

Lexical Entailment: Entailment is determined based on lexical concepts. A hypothesis is valid if its lexical components are true [1]. It is based on a probabilistic model and does not consider syntax and semantics.

Machine Learning approaches: Linear classifiers, logistic regression, support vector machines are classifiers used to train and learn from a dataset of text hypothesis pairs. It is a feature-based approach using similarity measures on words, stems, POS tags, chunk tags, negation, length ratio, of best partial match [8].

Graph based approaches: Text and hypothesis can be represented as directed graphs (dependency graphs), nodes representing words or phrases, and edges representing the relation between nodes [9]. Entailment is determined in these graphs using a matching cost based on vertex substitution and path substitution.

Deep learning approaches: The entailment recognition attempts in English from 2005 to 2015 are either rule-based or feature-based machine learning approaches. With the introduction of the SNLI dataset in 2015, a large dataset has enabled deep techniques for sentence representation using LSTM (Long Short Term Memory), CNN (Convolutional Neural Network) [10], BERT [11], and other transformer models and classification through deep neural networks [12]. Textual entailment is also used for fake news detection [13].

a) Datasets for Textual Entailment in English: The current works are mainly carried out in datasets, namely, RTE, SNLI, MNLI, and XNLI and in legal texts [14]. The collection of RTE datasets with their specifications are mentioned in the Table I.

TABLE I. RTE Datasets [15]

Dataset	size	Specification
RTE1	1367	manually collected pairs
RTE2	800	more realistic examples
RTE3	800	more longer texts
RTE4	1000	3 way classification (Entailment, Contradiction and Unknown)
RTE5	600	unedited texts
RTE6	15955	221 hypothesis
RTE7	21420	longer texts.

Other NLI datasets are SNLI (Stanford Natural Language Inference) dataset which is a collection of 570k English sentence pairs collected using Amazon mechanical trunk [3], and MNLI, Multi-Genre Natural Language Inference dataset is a collection of 433k sentence pairs from multiple genres [4].

b) Other languages: Entailment recognition in Japanese, Simplified Chinese, and Traditional Chinese language is attempted with RITE (Recognizing Inference in Texts) dataset [16], which has forward entailment, reverse entailment, bidirectional entailment, contradiction, independence as different classes for the Chinese sentence pairs. Surface textual features, lexical-semantic feature, syntactic feature, linguistic feature are used for classification using an SVM model [17].

Italian dataset is used in EVALITA campaign 2009 to recognize entailments in Italian text pairs [18]. Arabic dataset for textual entailment is detailed in [19]. Traditional features and distributed representations are used for recognizing textual entailment in Arabic [20]. Cross-lingual natural language inference dataset (XNLI) derives its collection from MNLI dataset and contains translated pairs in 15 languages, namely French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic,

Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu, out of which Hindi, Swahili, and Urdu are Indian languages [21]. Textual entailment for Indo-Aryan languages like Hindi is important to the language community of Northern parts of India. In this attempt we focus on Malayalam language from the Dravidian family. The Dravidian languages are mostly spoken in southern parts of India and has very minimal contributions when considering inferences. Attempts to different families of languages helps to gather significant contributions which are specific to those languages or language family and generic to all languages.

III. Challenges and Contributions

The Malayalam language is a South Indian Dravidian language, which has minimal works for textual entailment. The automatic and manual translation of SNLI pairs with linguistic corrections by experts forms the basis for the MaNLI (Malayalam Natural Language Inference) dataset. Prior work in Malayalam textual entailment reports the use of different embedding techniques, namely, Doc2Vec (paragraph vector), fastText, BERT(Bidirectional Encoder Representations from Transformers) and LASER(Language Agnostic SEntence Representations) for embedding sentence pairs for classification through Densenet [22]. Another attempt use siamese networks for binary classification of inference in texts [23]. The accuracy measure in Table II shows that LASER embedding based classification achieves the best results.

TABLE II. Performance of Different Embedding Methods for NLI in Malayalam

Embedding method	Binary	Multiclass
Doc2Vec	0.58	0.49
fastText	0.68	0.52
BERT	0.66	0.5
LASER	0.77	0.64

A. MaNLI Dataset

The development of language resources for Malayalam is in a progressing stage by different organizations and individual contributors. The Malayalam Natural Language Inference dataset is a dataset developed for natural language inference in the Malayalam language. It is created by manual and machine translation of text hypothesis pairs from the SNLI (Standford Natural Language Inference) dataset. Certain incorrect translations were corrected through manual efforts. Olam dictionary [24] is also used to get common substitutes for the English words. The details of the dataset are in Table III.

TABLE III. Dataset Statistics

MaNLI dataset	
Total sentence pairs	12000
Entailment pairs	4026
Contradiction pairs	3963
Neutral pairs	4011
Unique words	16194
Avg. premise sentence length	9.17
Avg.hypothesis sentence length	5.04

The dataset is created because an adequately annotated and linguistically correct entailment dataset is unavailable in this language. Hence, the translation method with linguistic corrections from language experts is adopted as one method to produce this dataset. This method involves less time and cost than creating an entirely new dataset that requires more time and human involvement to create sentence pairs and annotations.

The MaNLI dataset [22] [25] is a collection of 12K text-hypothesis pairs classified into entailment, contradiction, and neutral. Translations are done in such a way that the semantic content is maintained the same. Hence the annotated class labels are reused. It has been manually verified by linguists from the Thunchath Ezhuthachan Malayalam University, Kerala. The sentence length distribution for text and hypothesis sentences from the corpus is shown in Fig. 1. The premise sentences have word length between 5 and 15 whereas the hypothesis word length varies between 5 and 10 for most cases.

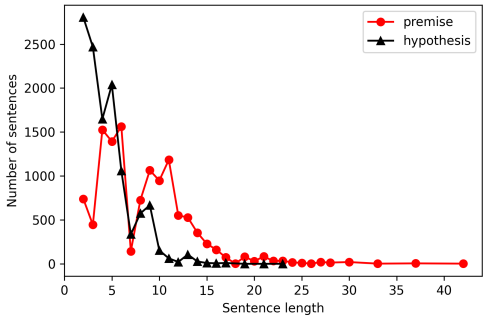


Fig. 1. Sentence Length Distribution

IV. Proposed Method

Textual entailment or natural language inference in English is attempted using machine learning and also deep learning approaches. But feature based machine learning approaches are not reported for the Malayalam language. In this work, we aim to develop systems for the Malayalam language using feature-based machine learning methods, which is essential to understand any classification problem. Also, comparison of feature-based models with deep learning methods became more feasible and realistic.

The design of the proposed work is shown in Fig. 2. Input pairs of text and hypothesis are preprocessed, and various lexical, semantic, and set-based features are extracted. The machine learning module classifies the text hypothesis pairs based on the extracted features using ML algorithms, such as Logistic regression, Support Vector Machine, Decision tree, Random Forest, Multinomial Naive Bayes and Adaboost.

A. Preprocessing

The sentence pairs are split into tokens, and prefixes and suffixes are removed in the preprocessing stage through tokenization and stemming. Tokenization is the process by which the words in the sentences are split into individual units called tokens for processing. The splitting is done using space as a separator. Stemming removes affixes from words.

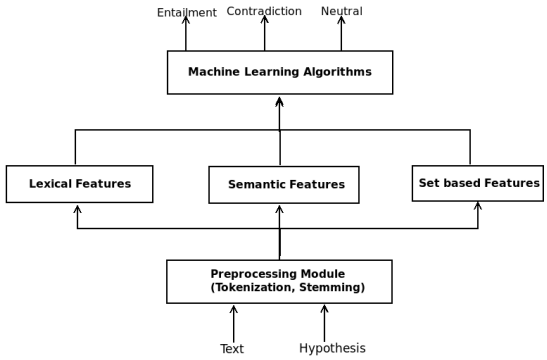


Fig. 2. Design of Textual Entailment Recognition System using Feature based Machine Learning.

For example, the word 'flowers' can have its stem word as 'flower', removing the suffix 's'. For the Malayalam language, libindic stemmer [26] available online is used. It is a rule-based stemmer using iterative suffix stripping to handle inflectional words.

B. Feature based Classification

This section details the different features used for the entailment classification. The features fall into different categories, namely lexical features, semantic features, and set-based features.

Lexical features: Lexical features are word or surface-level features that deal with the overlap of words. The different lexical features used are:

1) Unigram match: Overlap score of unigrams in text and hypothesis is computed. The unigram match score for text is defined as the number of unigram overlaps by the total number of unigrams in text. The Unigram match score for the hypothesis is defined as the number of unigram overlaps by the total number of unigrams in the hypothesis.

2) Bigram match: It is the number of bigram overlap divided by the number of bigrams in hypothesis.

3) Longest Common Subsequence: The LCS match is calculated as the length of the longest common subsequence/length of hypothesis.

4) Skip gram match: Skip-gram is a combination of n words in the sentence with few gaps. Skip grams with degree 2 and skip distance 1 are found for text and hypothesis. These skip grams matched count is divided by the number of skip grams in the hypothesis.

5) Length features: This consists of different length measures, such as $|B-A|$, $|A\&B|$, $(|B|-|A|)/|A|$, $(|A|-|B|)/|B|$, $|A\&B|/|B|$ where A is the text and B is the hypothesis.

Semantic features: Semantic features deal with the semantics of the sentences. For this, we have used word vector representation and term frequency-inverse document frequency (TF-IDF) of sentences.

1) Word embedding similarity: Word vectors from Word2Vec [27] are used to represent the words. Summation of word vectors of a sentence (text/hypothesis) is computed, and cosine between the two gives a similarity feature value.
2) TF-IDF similarity: Term Frequency -Inverse Document Frequency (TF-IDF) is a numerical statistic that evaluates the importance of a word in a collection. It is the product of term frequency and inverse document frequency. Text and hypothesis are represented using TF-IDF representation and cosine similarity is computed between the two.

Set/Distance based measures: Set/Distance based measures are the different types of similarities using counts for set-based unions and intersections. The various set-based similarities are:

1) Dice similarity: It measures the spatial overlap between two sentence pairs.

$$Dice(X,Y) = \frac{2\,\|X \cap Y\|}{\|X\| + \|Y\|} \tag{1}$$

If X and Y are similar, Dice coefficient will be 1 and otherwise 0.

2) Cosine similarity: It measures the cosine of the angle between the two sentences.

$$Cosine(X,Y) = \frac{\|X \cap Y\|}{\sqrt{\|X\|.\|Y\|}} \tag{2}$$

3) Levenstein similarity: It measures the minimum number of insertions, deletions and substitutions required to transform one word to another.

4) NeedleWunsch similarity: It is a sequence alignment based similarity measure. It measure global alignment score by finding the no of edits required which is calculated from the alignment matrix.

5) Smith Watermann similarity: It is a dynamic programming method that uses local alignment as a metric to measure similarity. The alignment matrix is created with no negatives and the scores are calculated.

6) Jaro Winkler similarity: It is also a string metric that measures the edit distance between two sequences from beginning to a set of prefix length.

$$sim = sim_{\mathrm{j}} + lp(1 - sim_{\mathrm{j}}) \tag{3}$$

where sim_{j} is the Jaro similarity between strings s_1 and s_2, l is the prefix length, p =0.1 (constant scaling factor).

$$sim_{\mathrm{j}} = \begin{cases} 0 \; if \; m = 0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) \end{cases} \tag{4}$$

where $|s|$ is the string length, m = no of matching characters, t = no of transpositions.

7) Jaccard similarity: This metric has the ratio of similarity and dissimilarity of sample sequences.

$$Jaccard(X,Y) = \frac{\|X \cap Y\|}{\|X \cup Y\|} \quad (5)$$

C. Machine Learning Approaches

Inference in the Malayalam language is considered as binary and multiclass classification. Binary classes are entailment and contradiction. Multiclass includes entailment, contradiction, and neutral. The following machine learning algorithms are used to evaluate the performance.

1) Logistic Regression: Logistic regression has dependent variable in two classes. With two classes x1 and x2 and the binary response variable Y (p= P(Y=1)),

$$logistic\ regression,\ l = log_b \frac{p}{1-p} \quad (6)$$

Binary classification is done with liblinear solver and class weight is balanced. Multinomial logistic regression is used to predict the different possible outcomes of a categorically distributed dependent variable. The classifier with multinomial class weights and lbfgs solver is used for multiclass classification.

2) Support Vector Machine: SVM maps the training examples to points in n-dimensional space. For binary classification, it maps into a 2-D plane separated by a line, and samples are mapped into either of the side of the plane. For multiclass classification, the samples are separated into different categories by a hyperplane.

3) Random Forest: It is an ensemble learning method which constructs many decision trees at training. For classification task, output class is the class selected by majority of the trees.

4) Decision Tree: It has a predictive modeling approach, start of the tree has different observations, that it traverse through the branches and ends in leaf nodes belonging to the target category for the sentence pair.

5) MultinomialNB: It is a Naive Bayes classifier for multi class classification. The feature vector consists of frequencies or integer counts. It is based on the Bayes' theorem stated below: P(c |x) = P(x |c) * P(c) / P(x) where c is a class and x is the sample instance that is to be classified.

6) AdaBoostClassifier: Also called adaptive boosting, it consists of weak classifiers in which one of the classifier is used to train on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

V. Experimental Settings and Evaluation

Implementations are done in Spyder integrated environment. The libraries used are Libindic stemmer for stemming, NLTK toolkit for extracting bigrams, text distance library for evaluating the distance between two or more sequences, and Scikit Learn for machine learning algorithms and classification reports. Grid searchCV is used for SVM classification. Table IV shows the specific settings applied in Scikit Learn based classifiers.

TABLE IV. Experimental Settings for LR, SVM and Random Forest

Model	Settings
Logistic Regression	solver= liblinear, class weight=balanced (Binary)
Logistic Regression	solver= lbfgs, multiclass=multinomial (Multiclass)
Support Vector Machine	kernel: ovr, rbf, C:(1, 10), gamma: (1, 0.1, 0.01, 0.001, 0.0001)
Random Forest	no of estimators=100, max depth=5

We have used different combinations of the feature set to arrive at the results. The different feature set configurations are in Table V.

TABLE V. Different Feature Set Configurations based on Combination of Features.

Feature set	Features
F1	Lexical (L)
F2	Semantic (S)
F3	Distance (D)
F4	Lexical, Semantic (L,S)
F5	Lexical, Distance (L,D)
F6	Semantic, Distance (S,D)
F7	Lexical, Semantic, Distance (L,S,D)

Evaluation Metrics The classification performance is evaluated using the Scikit-Learn classification metrics namely accuracy, precision, recall and F1-score.

- Accuracy: Accuracy is defined as the ratio of number of correct predictions to total predictions. Accuracy $=(tp+tn)/(tp+fp+fn+tn)$
- Precision: Precision is defined as the ability of the classifier not to misclassify samples (label negative sample as positive). Precision $= tp/(tp+fp)$
- Recall: Recall is defined as the ability of the classifier to find all positive samples. Recall $= tp/(tp+fn)$
- F1-score: F1-score is the harmonic mean of precision and recall. F1-score $= 2.precision.recall/(precision+recall)$

where tp is true positive, fp is false positive, tn is true negative and fn is false negative.

TABLE VI. BINARY CLASSIFICATION RESULTS IN TERMS OF WEIGHTED AVERAGE ACCURACY, PRECISION, RECALL, F1-SCORE AND SUPPORT

Model	Accuracy	Precision	Recall	F1-score	Support
LR	0.66	0.66	0.66	0.66	1598
SVM	0.67	0.67	0.67	0.67	1598
RF	0.67	0.67	0.67	0.67	1598
DT	0.66	0.66	0.66	0.65	1598
MNB	0.62	0.62	0.62	0.61	1598
AdaBoost(AB)	0.66	0.66	0.66	0.66	1598

TABLE VII. MULTICLASS CLASSIFICATION RESULTS IN TERMS OF WEIGHTED AVERAGE ACCURACY, PRECISION, RECALL, F1-SCORE AND SUPPORT

Model	Accuracy	Precision	Recall	F1-score	Support
LR	0.48	0.48	0.48	0.48	2400
SVM	0.50	0.50	0.50	0.50	2400
RF	0.49	0.49	0.49	0.48	2400
DT	0.46	0.46	0.46	0.46	2400
MNB	0.42	0.42	0.42	0.42	2400
AdaBoost(AB)	0.49	0.49	0.49	0.49	2400

The results of the classification evaluated in terms of accuracy, precision, recall, and F1-score is shown in Table VI with the whole 7989 pairs for binary classification and Table VII with 12k pairs for multiclass classification with the feature set configuration F7 having all the features. The train test split is 80:20. The performance of the rest of the feature sets (F1 to F6) is low compared to F7, hence we selected the feature set F7 for our study and comparisons. The performance of other feature sets is detailed in Section VI-B. From Tables VI and VII, it can be inferred that SVM, random forest and AdaBoost better classifies the Malayalam texts into entailment, contradiction and neutral classes.

We have evaluated our system with an increasing size of the data ranging from 2000 to 12000. The variation in the performance is shown in Fig. 3 for binary classification. The plot for mutliclass classification is shown in Fig. 4.

VI. RESULTS AND DISCUSSION

A. Effect of Increasing Size of Dataset

This section discusses the difference in the performance of deep learning and feature-based machine learning classification for binary and 3-way classification. As the size of the dataset increases from 2000 to 12k, there is a reduction in performance of feature-based classification. The features selected may be suitable for a few samples, but they can be misleading for other samples. Hence the model is not able to generalize with the samples.

LASER-based approach [22]: In the case of deep learning approaches with embedding that captures the context and places the sentences in semantic space, the model can generalize in a much more efficient manner. Prior work on entailment classification using LASER based sentence embedding has a BiLSTM encoder trained for 93 languages and includes Malayalam also. With character and word level representations, it produces sentence embeddings which are mapped in a semantic space. A feed forward neural network having sigmoid/softmax activations classifies the dataset into binary/ 3-class. It is more generic approach and the size of the dataset is immaterial when using a pretrained model.

In Fig. 3, and Fig. 4, the notation 'LS' denotes the LASER-based approach using deep learning approach, and the rest are the machine learning feature-based methods. From the figure, it can be inferred that when the dataset size is around 2000, both machine learning and deep learning approaches perform similar classifications. As and when the data is increased, deep learning based methods become more suitable, and it is observed through the comparison with this feature-based machine learning implementation. It also supports the fact that earlier works in English with RTE datasets used feature based approaches.

With 2000 samples of data, we have obtained good results with feature based classification. As the sample size increases, deep learning methods became more efficient in classification supporting the related works with SNLI dataset. This work adds to the literature for Malayalam entailment or inference tasks as a baseline for machine learning based on the feature set approach, which is novel with respect to this language. As the dataset is generic in nature, the distinguishing characteristic of features becomes low, and this can lead to poor classification on large datasets. Thus the performance of feature-based classification is limited in terms of features that generalize well with datasets of high semantic variability. Hence, the rise in performance of deep learning approaches hints that these are methods that can be adopted from small to large datasets.

B. Ablation Study

The ablation study for this work includes the performance of different features contributing to the classification of inferences in text in the Malayalam language. With the set of features, namely, lexical, semantic, and distance measures, we have studied the performance of different feature set combinations, and the results are discussed here.

TABLE VIII. F1-SCORE FOR DIFFERENT FEATURE SET COMBINATIONS WITH DIFFERENT CLASSIFIERS.

	F1	F2	F3	F4	F5	F6	F7
LR	0.47	0.38	0.43	**0.48**	**0.48**	0.43	**0.48**
SVM	0.48	0.37	0.43	0.48	0.49	0.44	**0.50**
RF	**0.49**	0.39	0.43	**0.49**	0.48	0.43	0.48
NN	0.37	0.15	0.15	0.43	**0.49**	0.15	0.47
MNB	0.40	0.15	0.28	0.41	0.41	0.31	**0.42**
DT	**0.47**	0.36	0.41	**0.47**	0.46	0.38	0.46
AdB	0.48	0.38	0.42	0.48	0.48	0.43	**0.49**

TABLE IX. MODEL SELECTION

Feature set	Performance (#classifiers with max F1-score/#classifiers)
F1	0.29
F2	0
F3	0
F4	0.43
F5	0.29
F6	0
F7	**0.57**

The chosen setting for the experimental results combines lexical, semantic, and distance measures (F7). Also, we have studied the model performance with only lexical (F1), semantic (F2), distance-based (F3), lexical and semantic (F4), lexical and distance (F5), and semantic and distance-based (F6). Based on Table VIII, the feature set that performs good on a

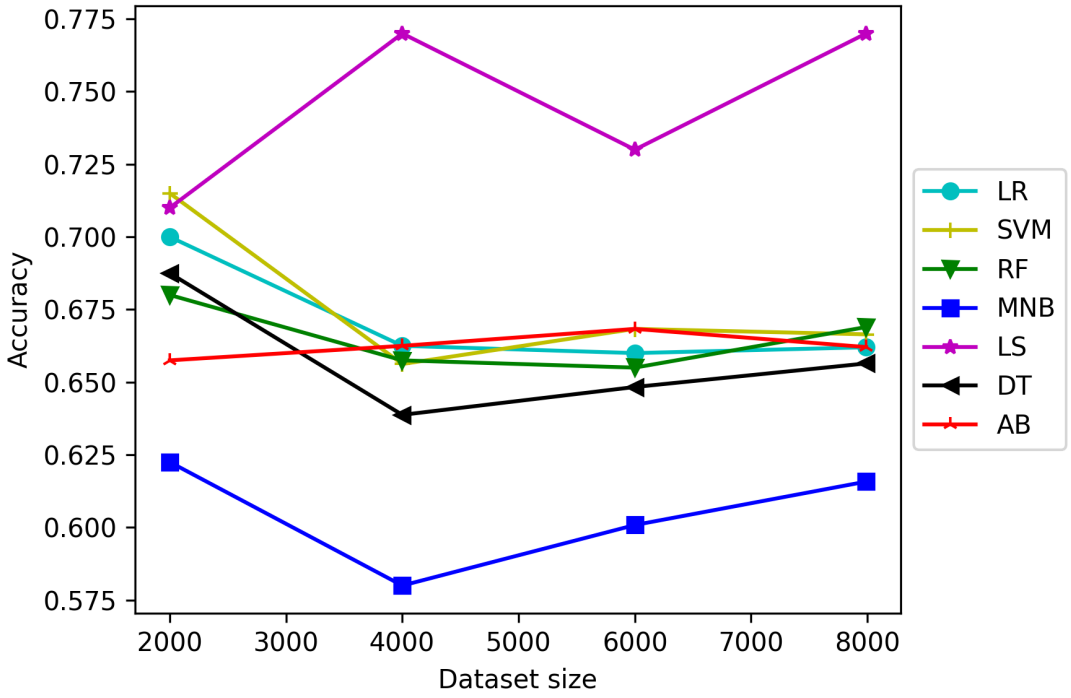


Fig. 3. Accuracy based Comparison (ML vs DL) Plot for Binary Classification, ML Methods are LR: Logistic Regression, SVM: Support Vector Machine, RF: Random Forest, MNB: Multinomial Naive Bayes, DT: Decision Tree, AB: Adaptive Boosting, DL method is LS: LASER based classifier.

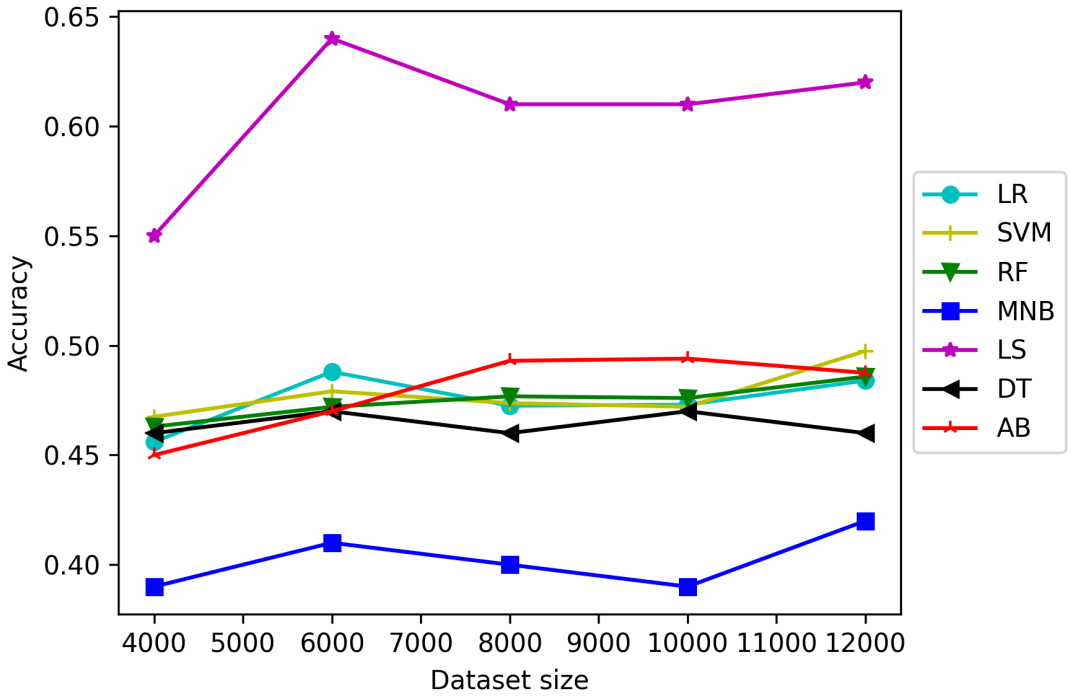


Fig. 4. Accuracy based Comparison (ML vs DL) Plot for Multiclass Classification, ML Methods are LR: Logistic Regression, SVM: Support Vector Machine, RF: Random Forest, MNB: Multinomial Naive Bayes, DT: Decision Tree, ADB: Adaptive Boosting, DL method is LS: LASER based Classifier.

majority of classifiers is chosen for analysis and comparison. The feature set performance is evaluated as in Table IX. The feature set performance is evaluated as the ratio of the number of classifiers with maximum F1-score to the total number of

classifiers. This justifies the selection of feature set F7, having maximum performance for experimental evaluations.

VII. Conclusion And Future Work

In this work, textual entailment is recognized for the Malayalam language with a feature-based approach. A set of classifiers are used to evaluate the performance accuracy. The best feature set model is chosen based on the F1-score measures. It is the first feature based attempt in this language for textual entailment recognition. This method also helped us understand the significant performance of deep learning methods, which is evident in the comparison. Thus this work on feature-based textual entailment recognition for the Malayalam language is substantial to the language resources community. The work is also essential and useful in identifying inferences in Malayalam texts for various language processing and social networking applications. Future work can include deep learning models to recognize entailment and these systems can be used in language processing applications.

Acknowledgment

The authors would like to thank the Department of Computer Science, CUSAT for the support extended in carrying out this research work.

References

[1] O. Glickman and I. Dagan, "A probabilistic setting and lexical coocurrence model for textual entailment," in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 2005, pp. 43–48.

[2] S. Ghuge and A. Bhattacharya, "Survey in textual entailment," *Center for Indian Language Technology, retrieved on April*, 2014.

[3] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[4] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1112–1122.

[5] P. Pakray, S. Bandyopadhyay, and A. Gelbukh, "Textual entailment using lexical and syntactic similarity," *International Journal of Artificial Intelligence and Applications*, vol. 2, no. 1, pp. 43–58, 2011.

[6] P. Pakray, S. Neogi, P. Bhaskar, S. Poria, S. Bandyopadhyay, and A. F. Gelbukh, "A textual entailment system using anaphora resolution." in *TAC*, 2011.

[7] T. Saikh, S. K. Naskar, C. Giri, and S. Bandyopadhyay, "Textual entailment using different similarity metrics," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 491–501.

[8] P. Malakasiotis and I. Androutsopoulos, "Learning textual entailment using svms and string similarity measures," in *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, 2007, pp. 42–47.

[9] R. Basak, S. K. Naskar, P. Pakray, and A. Gelbukh, "Recognizing textual entailment by soft dependency tree matching," *Computación y Sistemas*, vol. 19, no. 4, pp. 685–700, 2015.

[10] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[11] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[12] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[13] P. K. Rath and R. Basak, "Automatic detection of fake news using textual entailment recognition," in *2020 IEEE 17th India Council International Conference (INDICON)*, 2020, pp. 1–6.

[14] J. Rabelo, M.-Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, and K. Satoh, "Coliee 2020: methods for legal document retrieval and entailment," in *JSAI International Symposium on Artificial Intelligence*. Springer, 2020, pp. 196–210.

[15] I. Dagan, B. Dolan, B. Magnini, and D. Roth, "Recognizing textual entailment: Rational, evaluation and approaches–erratum," *Natural Language Engineering*, vol. 16, no. 1, pp. 105–105, 2010.

[16] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda, "Overview of ntcir-9 rite: Recognizing inference in text." in *Ntcir*. Citeseer, 2011.

[17] M. Liu, Y. Guo, and L. Nie, "Recognizing entailment in chinese texts with feature combination," in *2015 International Conference on Asian Language Processing (IALP)*. IEEE, 2015, pp. 82–85.

[18] J. Bos, F. M. Zanzotto, and M. Pennacchiotti, "Textual entailment at evalita 2009," *Proceedings of EVALITA*, vol. 2009, no. 6.4, p. 2, 2009.

[19] M. Alabbas, "A dataset for arabic textual entailment," in *Proceedings of the Student Research Workshop associated with RANLP 2013*, 2013, pp. 7–13.

[20] N. Almarwani and M. Diab, "Arabic textual entailment with word embeddings," in *Proceedings of the third arabic natural language processing workshop*, 2017, pp. 185–190.

[21] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," in *EMNLP*, 2018.

[22] S. Renjit and S. Idicula, "Natural language inference for malayalam language using language agnostic sentence representation," *PeerJ Computer Science*, vol. 7, p. e508, 2021.

[23] S. Renjit and S. M. Idicula, "Siamese networks for inference in Malayalam language texts," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sep. 2021, pp. 1167–1173. [Online]. Available: https://aclanthology.org/2021.ranlp-main.131

[24] K. Nadh, "Olam dictionary," https://olam.in/, 2010, accessed: 2021-11-16.

[25] S. Renjit, S. & Idicula, "Manli dataset," https://github.com/SaraRenG/MaNLI-data, 2020, accessed: 2021-11-16.

[26] S. Thottingal, "Libindic stemmer," https://github.com/libindic/indicstemmer, 2018, accessed: 2021-11-16.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

Towards Linguistic-based Evaluation System of Cloud Software as a Service (SaaS) Provider

Mohammed Abdulaziz Ikram[1], Ryan Alturki[2], Farookh K. Hussain[3]
Department of Computer Science, University College in Al-Jamoum,
Umm Al-Qura University, Saudi Arabia[1]
Department of Information Science, College of Computer and Information Systems,
Umm Al-Qura University, Makkah, Saudi Arabia[2]
School of Software, Faculty of Engineering and Information Technology,
The University of Technology Sydney, Ultimo, NSW
Australia[3]

***Abstract*—Cloud Software as a Service (SaaS) is a type of delivering software application by using Cloud computing Infrastructure services. Cloud SaaS used the global Internet connection to offer its services to the client consumers. The selection of Cloud SaaS provider is based on the evaluation mechanism that the Cloud SaaS consumer manage before making the service contract. In this paper, the linguistic-based evaluation of Cloud SaaS quality attributes has been proposed to help the consumer to assess the service for optimal service selection. Our proposed approach has been developed by the combinations of fuzzy logic and TOPSIS MCDM methods. The proposed approach helps the Cloud SaaS consumer to select the optimal service Cloud SaaS service provider. The case study has been proposed in order to demonstrate the proposed approach.**

***Keywords*—Cloud services; software as a service (SaaS); evaluation system; quality of experience (QoE); fuzzy logic; TOPSIS**

I. Introduction

Cloud SaaS is one of three types of services that Cloud Computing provides. It is a kind of software component that runs on top of a platform as a Service (PaaS) that in turns works on top of the Infrastructure as a Service [1]. Cloud SaaS is accessed and utilized via the global network infrastructure (i.e. the Internet) rather than installing software on computer client machine (i.e. on-premises service) [2]. Cloud SaaS runs on vendor's datacenter Cloud without installing it into the end-user machine. The user has full administrative control of functions such as insert, edit, delete and etc. However, the Cloud SaaS vendor has the responsibility for care of customer data and upgrading and modifying without any associated burden to the user. In recent years, there has been notable growth in the adoption of Cloud SaaS as it can reduce the cost of building software applications and services, especially for small and medium enterprises (SMEs) [3]. The Cloud SaaS has therefore become the leading and fastest growing software development paradigm comparing with licensed on-premises software service which may classified as an old fashion of delivering software [4], [5].

In recent years, research on Cloud SaaS selection has gained much attention by developing the evaluation system to assess the selected service provider [6], [7], [8], [9], [10]. The majority of existing works have developed the evaluation system based on pairwise comparing each quality evaluation attribute in order to make a decision and select the optimal service. In addition, the existing works consider the quantitative evaluation mechanisms to make a decision for selecting the optimal service. Therefore, evaluation of the Cloud SaaS provider based on subjective quality attributes needs to be addressed for further studies.

In this study, a new technique that improves the evaluation of Cloud SaaS provider is suggested. The proposed approach is existing works that proposed in our previous works for ranking the services based on the consumer's preferences [11], [12], [13]. Our proposed approach introduces the main quality evaluation attributes based on asking expert people of selection Cloud SaaS provider as well as reviewing the literature on the field of Cloud SaaS selection. The fuzzy logic is developed to interpret the linguistic terms in order to evaluate the quality evaluation attributes. Moreover, the multi-criteria decision making (MCDM) approach has been utilized by TOPSIS mathematical method in order to select the optimal service after the evaluation process. In the end, a case study is developed for demonstration purpose of Linguistic-based evaluation system proposed approach.

The remainder of the paper is organized as follows: Section 2 discusses outlines the literature review on the Cloud SaaS Evaluation system, Section 3 describes the proposed evaluation system framework. The case study is presented in Section 4; Section 5 concludes the paper.

II. Literature Review

Evaluation of Cloud SaaS provider aims to help a Cloud SaaS consumer to assess selected Cloud SaaS providers for evaluation purposes [9], [10]. As a result, there is an improved selection process with greater satisfaction in Cloud SaaS provider choice. The majority of Cloud SaaS providers offer a month trial version with limited capabilities to help a service consumer to test and evaluate their potential services to make an educated decision on whether or not to continue using the service or not. Therefore, the evaluation system is based on the consumer perspective, which is known as quantifying a service consumer's experiences. A few techniques [6], [7], [8], [9], [10] proposed on evaluation methods for the Cloud SaaS provider.

Godse and Mulik (2009) proposed evaluation method for Cloud SaaS provider using the analytic hierarchy process (AHP) approach in order to select an optimal service. They also consider some quality attributes for evaluation purposes such as service functionality, architecture, usability, vendor, reputation and cost [6]. In addition, and based the same evaluation mechainsm, Boussoualim and Aklouf (2015) proposed pairwise comparison for weighting the quality attributes as well as an evaluation service provider, and prior to decision making regarding a service. The authors were also considering the functional category of Cloud SaaS providers and non-functional quality attributes for the evaluation process. The main non-functional quality attributes considered in their work as follows: reputation, cost, usability, structure, configurability and personalization. The authors finally demonstrated their approach by a case study [9]. However, from existing approaches developed the evaluation system by an AHP method is only suitable with a small number of alternatives.

Reixa et al. (2012) proposed an evaluation approach based on aggregation function based on a Cloud SaaS consumer perspective with also considering the opinion of an expert person. The highest score indicates that the service has many of the appropriate features. The authors provide factors that a service consumer should consider when evaluating a Cloud SaaS provider. The factors generated by extensive studies using the expert group report "European commission" that focus and analyze the main characteristics of cloud services. These factors as follows: suitability, economic value, control, usability, reliability and security. The author also provides a case study for selecting only three office cloud services: Zoho Docs, google docs and Microsoft office365 [7].

Jagli et al. (2016) proposed an evaluation mechanism using a decision tree-data mining model. The authors investigate the main challenges for representing the qualities that a Cloud SaaS consumer are considering when evaluating a Cloud SaaS service provider. They concluded with the following criteria: pay-per-use, availability, reusability, scalability, data managed by provider and service customizability [8]. However, the main shortcoming of this work is the lack of demonstration of its use in a case study or experiment.

Naseer and Nazar (2016) proposed an evaluation of Cloud SaaS provider based on the analytic network process (ANP) to make a decision and then select the best service provider. The authors also investigated the main quality attributes for evaluating Cloud SaaS providers including usability, security, reliability, tangibility, responsiveness and empathy [10].

Therefore, we can summarize the main shortcomings of existing approaches as follows: (1) none of the existing approaches are considered an evaluation process using the consumer's experience perspective. (2) none of the existing approaches are used the evaluation selected the Cloud SaaS providers based on linguistic-based evaluation quality attributes. In order to address the above shortcomings, the evaluation mechanism based on the linguistic evaluation quality attributes is proposed. The system proposed is developed by combining the fuzzy logic with multi-criteria decision making.

III. Linguistic-Based Evaluation System Framework

The main contribution of this work is for evaluation Cloud SaaS providers based on subjectively quality attributes. In our previous work [11], [12], [13], the Cloud SaaS ranking system proposed to help Cloud SaaS consumer for ranking and sorting the service providers based on quantitative quality attributes requested by service consumer. Our proposed system consists of three main parts as shown in Fig. 1: (1) Quality of Experience (QoE) Service Repository -$QoESR$, (2) Consumer Evaluation Handler CEH and (3) Decision Maker System - DMS.All the components are worked together in order to make a decision to select an optimum service provider. Each of these elements are discussed in the following subsections.

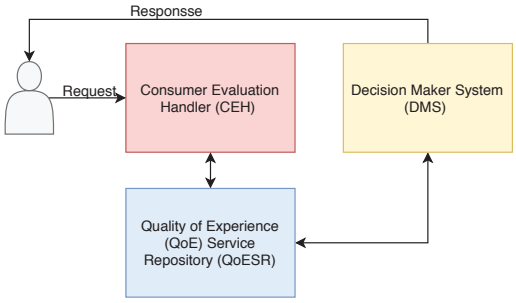


Fig. 1. Linguistic-based Evaluation System

A. Quality of Experience (QoE) Service Repository (QoESR)

This section details the service repository required to deal with the Cloud SaaS providers and the quality evaluation attributes. As was mentioned before, our proposed system considers the evaluation of service providers based on the consumer's experience. Therefore, the quality of experiences (QoEs) has been examined intensively to consider the most concerning factors when evaluate the service. However, due to the lack of a standard repository for the QoE for the Cloud SaaS service provider, we propose a new standard of QoE based on the views of an expert team and with consideration of the literature review in this section.

The QoEs have divided into three categories based on (1) service vendor, (2) service data and (3) service itself. Table I details of the quality of experiences (QoEs) that have been collected for our study.

1) **Vendor category** includes all factors affecting a Cloud SaaS consumer when selecting a service related to a Cloud SaaS vendor. The vendor category involves software support, vendor reputation and training.
    a) **Software support**: Cloud SaaS providers offer several support strategies to their service consumers, such as communication via chatbots, email and hotline telephone numbers in case of emergency. Cloud SaaS consumers can check how quickly and efficiently software vendors solve software conflicts and

TABLE I. EVALUATION QOE

QoE Factor	Evaluation Concerns
Software Support	Helpdesk support, software update and problem solving response.
Vendor Reputation	Service history, number of clients, service rating and social media rating.
Training	Training materials, videoed explanations and community support.
Security	Security certificates, HTTPS and trust certificate.
Recoverability	Supporting service backup.
Interoperability	Supporting the export of different kinds of files.
Usability	Easy-to-use and supporting Mobile App.
Integration	Supporting integration with other services.
Documentation	E-book, White papers and service report.
Offline support	Supporting offline service and synchronizes service automatically after reconnecting service.

usability problems and provide updates to fix bugs and improve security.

b) **Vendor reputation**: It is important that the customer considers the reputation of the service provider, taking into account the number of clients who have used the software and their satisfaction level. Another consideration is the brand value of the software vendor. Some consumers prefer to select an established software vendor which has a good reputation, however, a new software vendor may have a good reputation if their services have expanded and if they provide sufficient documentation on the Internet [6], [9].

c) **Training**: A service consumer should also consider whether a vendor provides training, such as workshops, tutorials, or webinars covering both local and international events and information on video channel platforms such as YouTube to help consumers understand the full range of features offered by the service.

2) **Data category** includes all the data-related factors of Cloud SaaS. The data category comprises security, recoverability and interoperability.

a) **Security**: This tends to be the main concern for most Cloud SaaS service consumers since SaaS is a multi-tenant service concept. Security is a process or set of actions to ensure data is protected from unauthorized persons or systems [7]. Security involves confidentiality, authenticity, integrity, and availability. A Cloud SaaS service consumer should check if a service is certified with security certifications such as SAS 70, SSAE 16, ISAE 3402, ISO/IEC 9126, ISO/IEC 27001, and ISO/IEC 27002 to ensure the service has an appropriate level of security [14].

b) **Recoverability**: Backup periodicity specifies how often backups are made, either in a continuous manner or at regular intervals. Recovery velocity specifies how quickly data can be recovered from the backup in case of application or infrastructure issues, including the backup service in terms of using on-premise devices or cloud storage options.

c) **Interoperability**: Data interoperability refers to the ability of data to be represented in different forms so that it can be accessible over different software platforms. A consumer needs to verify if a system can export their data in commonly used data forms. For example, if a service is selected by a hospital to manage the input of patient information, then the selected service must be able to export the data in various forms, such as PDF, XLS and CSV. This is an important feature in case a consumer wants to adopt another cloud service, or in the backup process, or even for further research or study purposes.

3) **Service Category** covers all the quality parameters involved with services such as the design of the windows screen and the integrity of the system. The service category comprises usability, integration, documentation and offline support.

a) **Usability** is a measure of the consumers satisfaction with a service and the quality of their experience in interacting with the Cloud SaaS [14]. A Cloud SaaS provider should ensure there is good user interface (UI) to enhance the users experience of service. Usability is evaluated in terms of the following parameters: (1) UI is intuitive and easy-to-use for frequently implemented tasks and has an attractive graphical windows design; (2) the availability of user manuals to enhance consumer understanding of the windows design and e-Learning modules; (3) support for mobile applications such as mobile phones and tablets [6].

b) **Integration** refers to the ability of software to communicate easily with other service platforms to share information and is generally connected to the application programming interface (API) which enables one service to be integrated with other services and systems [14]. A consumer should verify if the service API has the features to enable it to connect with other software platforms. For instance, if a consumer selects the ERP service and needs additional functional software such as a storage service, then the consumer should investigate whether the ERP service for the selected software is able to be integrated with other services, which in this case, is a storage service.

c) **Documentation** is important that the service consumer understands the functions of the service, and how the software operates in various stages. Documentation also includes different kinds of materials provided either by a service provider or service community or service consumers, such as E-Books, reports and white papers.

d) **Offline support** is an important factor due

to the continuous connectivity with the software server and measures whether the service supports the connectivity with the system in offline mode and synchronizes data once it reconnects to the Internet.

B. Consumer Evaluation Handler (CEH)

The main objective of CEH is to register the consumers feedback on the quality factors for each service provider and to assign weights to indicate the priority of the quality factor. The CEH deals with two important linguistic variables from a service consumer: (1) Quality evaluation (QEvaluation, ε), and (2) Quality priority (QPriority, ρ). The result after fuzzified is the quality evaluation score (QEvaluationScore, δ).

- **Quality Priority (QPriority)** (ρ_j)
  This is the first linguistic variable of our proposed system. It is used to assign weights to determine the importance of quality for the evaluation process. We propose seven linguistic values which are decomposed into triangular fuzzy numbers using the triangular fuzzy set shown in Fig. (2). Table II shows the seven linguistic terms which are used to measure the quality priority.

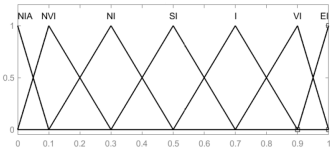


Fig. 2. Membership Function for Linguistic Values of QPriority

TABLE II. LINGUISTIC VARIABLES OF WEIGHTING QUALITIES WITH THEIR FUZZY NUMBERS

Linguistics term	Fuzzy number
Extremely Important (EI)	(0.9,1.0,1.0)
Very Important (VI)	(0.7,0.9,1.0)
Important (I)	(0.5,0.7,0.9)
Somewhat Important (SI)	(0.3,0.5,0.7)
Not Important (NI)	(0.1,0.3,0.5)
Not Very Important (NVI)	(0.0,0.1,0.3)
Not Important at All (NIA)	(0.0,0.0,0.1)

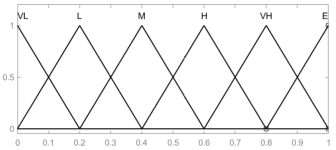


Fig. 3. Membership Function for Linguistic Values of QEvaluation

- **Quality Evaluation (QEvaluation, $\varepsilon_i j$)**
  The second linguistic variable in our proposed approach is the quality evaluation, $\varepsilon_i j$. A Cloud SaaS service consumer evaluates individual quality variable as linguistic values after trialling the service during the free test period. Six linguistic values are proposed for evaluation quality as shown in Table III. The membership functions of the linguistic values are shown in Fig. 3.

TABLE III. LINGUISTIC VARIABLES OF EVALUATING THE QUALITIES AND THEIR FUZZY NUMBER

Linguistic Values	Fuzzy numbers
Very low (VL)	(0.0.0.2)
Low (L)	(0,0.2,0.4)
Medium (M)	(0.2,0.4,0.6)
High (H)	(0.4,0.6,0.8)
Very high (VH)	(0.6,0.8,1)
Excellent (E)	(0.8,1,1)

- **Quality Evaluation Score (**$QEvaluationScore$**),** δ
  The final process before deciding and selecting a service is to aggregate the fuzzy weighting of the quality priority with the evaluation criteria to obtain the quality evaluation score. The quality evaluation score, δ, is the overall quality score of the individual quality of each service. This process is a metadata step to transmit all this information to the decision maker system to select the best service that matches the consumers requirements.

  Additionally, we need to aggregate these two fuzzy numbers together. To do this, the fuzzy number can be aggregated with different arithmetic equations [15] such as:

$$\tilde{a} \otimes \tilde{b} = (a_1, a_2, a_3) \otimes (b_1, b_2, b_3) = (a_1 \times b_1, a_2 \times b_2, a_3 \times b_3) \quad (1)$$

$$\tilde{a} \oplus \tilde{b} = (a_1, a_2, a_3) \oplus (b_1, b_2, b_3) = (a_1 + b_1, a_2 + b_2, a_3 + b_3) \quad (2)$$

$$\tilde{a} \ominus \tilde{b} = (a_1, a_2, a_3) \ominus (b_1, b_2, b_3) = (a_1 - b_1, a_2 - b_2, a_3 - b_3) \quad (3)$$

  Therefore, equation (1) is applied to aggregate these fuzzy numbers from the evaluation of services and the weighting criteria.

$$\begin{bmatrix} qoe_1 & qoe_2 & qoe_3 & ... & qoe_n \\ \tilde{\rho_1} \otimes \tilde{\varepsilon_1}1 & \tilde{\rho_2} \otimes \tilde{\varepsilon_1}2 & \tilde{\rho_3} \otimes \tilde{\varepsilon_1}3 & ... & \tilde{\rho_n} \otimes \tilde{\varepsilon_1}n \\ \tilde{\rho_1} \otimes \tilde{\varepsilon_2}1 & \tilde{\rho_2} \otimes \tilde{\varepsilon_2}2 & \tilde{\rho_3} \otimes \tilde{\varepsilon_2}3 & ... & \tilde{\rho_n} \otimes \tilde{\varepsilon_2}n \\ ... & ... & ... & ... & ... \\ \tilde{\rho_1} \otimes \tilde{\varepsilon_m}1 & \tilde{\rho_2} \otimes \tilde{\varepsilon_m}2 & \tilde{\rho_3} \otimes \tilde{\varepsilon_m}3 & ... & \tilde{\rho_n} \otimes \tilde{\varepsilon_m}n \end{bmatrix}$$

For the final step, the fuzzy interpreter will defuzzify the fuzzy number and transform all these values to a crisp number. To do this, there are different methods for the defuzzification of fuzzy numbers, such as the centre of gravity (CoG), First of Maximum (FOM), Last Of Maximum (LOM), COG (Center Of Gravity), Mean Of Maxima (MeOM), Weighted Fuzzy Mean (WFM), Quality Method (QM), Extended Quality Method (EQM) and Center Of Area (CoA) [16]. CoG has been used recently in different research to deal with linguistic terms [17]. Equation (4) is applied to transfer these values to

a crisp number.

$$d(\hat{A}) = \frac{1}{3}(a+b+c) \tag{4}$$

C. Decision Maker System (DMS)

The TOPSIS method is proposed to select the best Cloud SaaS after the evaluation under a fuzzy environment. The TOPSIS standard of a technique for order performance by similarity to the ideal solution was developed by Hwang Yoon [18]. The idea of TOPSIS is to select the best alternative based on the shortest distance to the positive ideal solution and the furthest distance from the negative ideal solution. The advantages of the TOPSIS method over the other MCDM approaches is that it is easy to develop with different programming language platforms. Moreover, TOPSIS can be used with many alternatives, which can be easily applied with $m \times n$ matrix. where m denotes the number of alternatives and n denotes the number of criteria. Other multi-criteria decision-making approaches such as AHP, ANP and ELECTRE are only suitable for use with a small number of alternatives due to the expense in processing and time where there are numerous alternatives.

The TOPSIS MCDM process comprises six steps to make a decision and selects the best alternative [18]. However, if we apply the TOPSIS method in our proposed Evaluation system to select the best service, we leave out some steps, such as the normalization process and calculating the weighting normalized matrix because these two steps have already been calculated by the fuzzy logic calculation proposed previously.

Accordingly, to apply the TOPSIS MCDM method into our proposed approach, four steps are needed to select the best Cloud SaaS service provider after the evaluation of the services by a cloud SaaS service consumer as follows:

Let us assume the QEvaluationScore matrix that will interact with the decision maker system is as follows:

$$\begin{bmatrix} qoe_1 & qoe_2 & qoe_3 & ... & qoe_n \\ \delta_1 1 & \delta_1 2 & \delta_1 3 & ... & \delta_1 n \\ \delta_2 1 & \delta_2 2 & \delta_2 3 & ... & \delta_2 n \\ ... & ... & ... & ... & ... \\ \delta_m 1 & \delta_m 2 & \delta_m 3 & ... & \delta_m n \end{bmatrix}$$

where $\delta_i j$ is the quality evaluation score after the defuzzification process for service i and quality j.

The steps involved in our method are as follows:

1) Calculate the positive ideal solutions and the negative ideal solutions. This step is used to select the service based on measuring the shortest distance to the positive ideal solution and the farthest distance to the negative ideal solution. The result of each calculation will be a vector size of $1 \times n$. where n denotes the number of qualities that are of concern for a service consumer. Equation (5) is used to calculate the positive ideal solution, and equation (6) is used to calculate the negative ideal solution.

$$A^* = v_1^*, v_2^*, ..., v_j^*, ..., v_n^* = \{(\max_i v_{ij} | j \in J_1), (\min_i v_{ij} | j \in J_2) | \, i = 1, ..., m\} \tag{5}$$

$$A^- = v_1^-, v_2^*, ..., v_j^-, ..., v_n^- = \{(\min_i v_{ij} | j \in J_1), (\max_i v_{ij} | j \in J_2) | \, i = 1, ..., m\} \tag{6}$$

2) Calculate the separation measures for each service from the positive and negative ideal solution. To do this, the TOPSIS method applies Euclidean distance to measure the distance to the positive and negative ideal solution. The result of the separation measures will be between [0,1]. Equation (7) is used to calculate the separation of each service from the positive ideal solution and equation (8) is is used to calculate the separation of each service from the negative ideal solution.

$$S_i^* = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^*)}, i = 1, ..., m \tag{7}$$

$$S_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)}, i = 1, ..., m \tag{8}$$

3) Calculate the relative closeness to the ideal solution for each cloud service using equation (9). The result of this step is between [0,1]. The service that has a higher number indicates better performance and that it is a good match with the consumer's requirements.

$$C_i^* = \frac{S_i^-}{S_i^* + S_i^-} \tag{9}$$

4) The last step is ranking and sorting the cloud services based on the index value C_i^j. The service that has a higher index value will be selected first which means it is very close to the consumer's evaluation of services. The result represents the evaluation score for each service which shows the position of each service.

IV. Case Study

To better illustrate our proposed evaluation approach, a case study is used to explain the process of evaluating services to make a final service selection using the linguistic-based evaluation technique.

Let us assume that a new business called Misbar is interested in adopting a Cloud SaaS solution using the CRM SaaS application. The expert team at Misbar understand the benefits of using cloud services instead of building their own software application, such as availability, elasticity and support for a mobile app. Therefore, six Cloud SaaS service provider were chosen for this purpose based on different QoS criteria such as service cost, availability and service rate.

For the final service selection, the Misbar team needs to evaluate their six services using the evaluation criteria

described in Table I. We assume all these selected services can be trialed and evaluated during the free test period offered by the service provider. As shown in Table IV, the Misbar team evaluates each service based on multiple QoE parameters. For instance, the first service provider evaluates software support as medium. This means this service has been evaluated as being medium in terms of supporting problem solving, fixing bugs, updating software and providing help desk support. Similarly, they evaluate the following criteria: vendor reputation, training, security, recoverability, interoperability, usability, documentation and offline support for all the selected services.

TABLE IV. LINGUISTIC VALUES TO EVALUATE THE SERVICES

Service Provider	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
SaaS provider 1	M	H	VH	M	VH	M	VH	M	L
SaaS provider 2	VL	L	M	M	VH	M	E	L	L
SaaS provider 3	VH	E	VH	VH	H	H	M	VH	H
SaaS provider 4	VL	H	H	H	M	M	M	L	L
SaaS provider 5	M	VH	L	M	H	H	M	L	M
SaaS provider 6	E	H	H	H	VH	M	VH	M	VH

After this, the Misbar team weights the priority of each quality attributes among other conflicting criteria to make a service selection. Table V shows the priority of the criteria provided by the Misbar team. As shown in this table, the Misbar team used linguistic values to weight the QoE parameters. For example, the Misbar team considers software support as very important, however, they consider the usability criteria as important.

TABLE V. LINGUISTIC WEIGHT CRITERIA

Weight criteria	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
Linguistic values	VI	I	EI	EI	VI	SI	I	NI	VI

Then, the Evaluation system transfers their linguistic values that describe the evaluation criteria into fuzzy numbers. Table VI illustrates the fuzzy numbers of the linguistic values for weighting the quality attributes.

TABLE VI. FUZZY NUMBERS FOR WEIGHTING CRITERIA

X	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
weight	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.9,1.0,1.0)	(0.9,1.0,1.0)	(0.7,0.9,1.0)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.1,0.3,0.5)	(0.7,0.9,1.0)

The Evaluation system then transfers the linguistic values of the evaluation of the quality attributes into fuzzy numbers. Table VII shows the fuzzy numbers of the evaluation criteria for all services.

TABLE VII. FUZZY NUMBERS FOR AN EVALUATION OF THE SERVICES

X	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
SaaS provider 1	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.0,0.2,0.4)
SaaS provider 2	(0.0,0.0,0.2)	(0.0,0.2,0.4)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.8,1.0,1.0)	(0.0,0.2,0.4)	(0.0,0.2,0.4)
SaaS provider 3	(0.6,0.8,1.0)	(0.8,1.0,1.0)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.4,0.6,0.8)
SaaS provider 4	(0.0,0.0,0.2)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.0,0.2,0.4)	(0.0,0.2,0.4)
SaaS provider 5	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.0,0.2,0.4)	(0.2,0.4,0.6)
SaaS provider 6	(0.8,1.0,1.0)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)

After collecting all the inputs from the Misbar team, the Linguistic-based Evaluation system is used to select the service which best matches the consumer evaluation for each service. Firstly, our system aggregates the fuzzy evaluation of the attributes (QEvaluation) with the fuzzy quality priority (QPriority) to obtain the quality evaluation score (QEvaluationScore). For example, regarding the first service, for the software support criteria, the fuzzy (QEvaluation) is (0.2,0.4,0.6) and the fuzzy (QPriority) is (0.7,0.9,1.0). Therefore, the aggregation of these two fuzzy numbers is $(0.2, 0.4, 0.6) \otimes (0.7, 0.9, 1.0)$ = $(0.2 \times 0.7, 0.4 \times 0.9, 0.6 \times 1.0)$. The final result is (0.14,0.36,0.6). Table VIII shows the evaluation score matrix for all cloud services with respect to the quality evaluation score.

TABLE VIII. FUZZY NUMBERS FOR EVALUATION MATRIX

X	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
Escore 1	(0.14, 0.36, 0.60)	(0.20, 0.42, 0.72)	(0.54, 0.80, 1.00)	(0.18, 0.40, 0.60)	(0.42, 0.72, 1.00)	(0.06, 0.20, 0.42)	(0.30, 0.56, 0.90)	(0.02, 0.12, 0.30)	(0.00, 0.18, 0.40)
Escore 2	(0.0, 0.0, 0.2)	(0.00, 0.14, 0.36)	(0.18, 0.40, 0.60)	(0.18, 0.40, 0.60)	(0.42, 0.72, 1.00)	(0.06, 0.20, 0.42)	(0.4, 0.7, 0.9)	(0.00, 0.06, 0.20)	(0.00, 0.18, 0.40)
Escore 3	(0.42, 0.72, 1.00)	(0.4, 0.7, 0.9)	(0.54, 0.80, 1.00)	(0.54, 0.80, 1.00)	(0.28, 0.54, 0.80)	(0.12, 0.30, 0.56)	(0.10, 0.28, 0.54)	(0.06, 0.24, 0.50)	(0.28 0.54 0.80)
Escore 4	(0.0, 0.0, 0.2)	(0.20, 0.42, 0.72)	(0.36, 0.60, 0.80)	(0.36, 0.60, 0.80)	(0.14, 0.36, 0.60)	(0.06, 0.20, 0.42)	(0.10, 0.28, 0.54)	(0.00, 0.06, 0.20)	(0.00, 0.18, 0.40)
Escore 5	(0.14, 0.36, 0.60)	(0.30, 0.56, 0.90)	(0.0, 0.2, 0.4)	(0.18, 0.40, 0.60)	(0.28, 0.54, 0.80)	(0.12, 0.30, 0.56)	(0.10, 0.28, 0.54)	(0.00, 0.06, 0.20)	(0.14 0.36 0.60)
Escore 6	(0.56, 0.90, 1.00)	(0.20, 0.42, 0.72)	(0.36, 0.60, 0.80)	(0.36, 0.60, 0.80)	(0.42, 0.72, 1.00)	(0.06, 0.20, 0.42)	(0.30, 0.56, 0.90)	(0.02, 0.12, 0.30)	(0.42 0.72 1.00)

Secondly, the fuzzy interpreter defuzzifies the fuzzy numbers for the (QEvaluationScore) to obtain the crisp number. To do this, the Center of Gravity (CoG) as proposed in Equation (4 is used to calculate the crisp number. For example, the fuzzy evaluation score for the first service for software support is (0.14,0.36,0.6). Therefore, to calculate the CoG is $\frac{1}{3}(0.14 + 0.36 + 0.6)$ with a result of 0.366666667. Table IX shows the crisp numbers of all the services.

TABLE IX. THE FINAL CRISP NUMBER OF ALL SERVICES

X	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
Escore 1	0.366666667	0.446666667	0.78	0.393333333	0.713333333	0.226666667	0.586666667	0.146666667	0.193333333
Escore 2	0.066666667	0.166666667	0.393333333	0.393333333	0.713333333	0.226666667	0.666666667	0.086666667	0.193333333
Escore 3	0.713333333	0.666666667	0.78	0.78	0.54	0.326666667	0.306666667	0.266666667	0.54
Escore 4	0.066666667	0.446666667	0.586666667	0.586666667	0.366666667	0.226666667	0.306666667	0.086666667	0.193333333
Escore 5	0.366666667	0.586666667	0.2	0.393333333	0.54	0.326666667	0.306666667	0.086666667	0.366666667
Escore 6	0.82	0.446666667	0.586666667	0.586666667	0.713333333	0.226666667	0.586666667	0.146666667	0.713333333

TOPSIS MCDM is a method which is used to assist decision making based on measuring the distance for each alternative based on the shortest distance to the positive ideal solution and furthest distance to the negative ideal solution. The positive ideal solution is the highest value for each criterion, whereas the negative ideal solution is the lowest value for each criterion. In the next step, our proposed system obtains the positive ideal solution and negative ideal solution to make a decision based on measuring the distance to these two vectors. Table X shows the positive ideal solution vector and negative ideal solution vector.

TABLE X. THE POSITIVE IDEAL SOLUTION AND NEGATIVE IDEAL SOLUTION

X	Software Support	Vendor reputation	Training	Security	Recoverability	Interoperability	Usability	Documentation	Offline Support
A	0.82	0.666666667	0.78	0.78	0.713333333	0.326666667	0.666666667	0.266666667	0.713333333
A-	0.066666667	0.166666667	0.2	0.393333333	0.366666667	0.226666667	0.306666667	0.086666667	0.193333333

After this, the TOPSIS method makes a decision by measuring the distance to the positive ideal solution (S^+) and the negative ideal solution (S^-) for each service using Euclidean distance. Table XI shows the distance to the positive ideal solution and the negative ideal solution for all services.

Before the last step, the TOPSIS method calculates the similarity to the positive ideal solution in order to obtain

TABLE XI. THE DISTANCE TO THE POSITIVE AND NEGATIVE IDEAL SOLUTIONS

SaaS SaaS Provider	S^+	S^-
Cloud SaaS 1	0.839417788	0.840819706
Cloud SaaS 2	1.195547294	0.535868972
Cloud SaaS 3	0.448404579	1.160478828
Cloud SaaS 4	1.119484008	0.515062024
Cloud SaaS 5	1.005009674	0.580076623
Cloud SaaS 6	0.392371706	1.142531303

the evaluation score for each service. Table XII shows the similarity of the positive ideal solution for all services. The values are between [0,1] where a value close to one indicates better service.

TABLE XII. SIMILARITY TO THE POSITIVE IDEAL SOLUTION

Cloud SaaS Provider	C^+
Cloud SaaS 1	0.500417179
Cloud SaaS 2	0.309497481
Cloud SaaS 3	0.721294547
Cloud SaaS 4	0.315110137
Cloud SaaS 5	0.365959017
Cloud SaaS 6	0.744367101

The final step is to rank the services based on the value of the similarity to the positive ideal solution. Table XIII shows the ranking of the services for the Misbar company. It can be seen that service provider 6 has the most similarity to Misbar's preferences at approximately 74.4 per cent. The service provider which is ranked last is service provider 2 with 30.9 per cent. Therefore, service provider 6 is the most suitable for selection based on Misbar's evaluation mechanisms.

TABLE XIII. RANKING OF THE SERVICES FOR SELECTION

SaaS Provider	C^+
Cloud SaaS 6	0.744367101
Cloud SaaS 3	0.721294547
Cloud SaaS 1	0.500417179
Cloud SaaS 5	0.365959017
Cloud SaaS 4	0.315110137
Cloud SaaS 2	0.309497481

V. CONCLUSION

This paper has been proposed an evaluation system based on the linguistic terms for selected quality attributes. The proposed work combines two mathematical models Fuzzy Logic plus the TOPSIS MCDM in order to make the final evaluation technique for each service provider. This work also gathers all evaluation attributes that the service consumer considers when selecting the Cloud SaaS services provider. The case study has been presented in this paper to demonstrate the proposed evaluation approach. Our approach helps the service consumer for selecting the optimal service provider among multiple similar services based on evaluated multiple attributes.

Our future dimension is to combine the quantitative and qualitative attributes to make final service provider selection.

REFERENCES

[1] W. Tsai, X. Bai, and Y. Huang, "Software-as-a-service (saas): perspectives and challenges," *Science China Information Sciences*, vol. 57, no. 5, pp. 1–15, 2014.

[2] R. Seethamraju, "Adoption of software as a service (saas) enterprise resource planning (erp) systems in small and medium sized enterprises (smes)," *Information systems frontiers*, vol. 17, no. 3, pp. 475–492, 2015.

[3] J. Repschlaeger, S. Wind, R. Zarnekow, and K. Turowski, "Selection criteria for software as a service: an explorative analysis of provider requirements," 2012.

[4] A. M. Alkalbani, A. M. Ghamry, F. K. Hussain, and O. K. Hussain, "Blue pages: Software as a service data set," in *2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)*, vol. 00, Nov. 2015, pp. 269–274. [Online]. Available: doi.ieeecomputersociety.org/10.1109/BWCCA.2015.83

[5] M. P. Papazoglou, "Service-oriented computing: concepts, characteristics and directions," in *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003.*, Dec 2003, pp. 3–12.

[6] M. Godse and S. Mulik, "An approach for selecting software-as-a-service (saas) product," in *Cloud Computing, 2009. CLOUD'09. IEEE International Conference on*. IEEE, 2009, pp. 155–158.

[7] M. Reixa, C. Costa, and M. Aparicio, "Cloud services evaluation framework," in *Proceedings of the Workshop on Open Source and Design of Communication*. ACM, 2012, pp. 61–69.

[8] D. Jagli, S. Purohit, and N. S. Chandra, "Evaluating service customizability of saas on the cloud computing environment," *Int. J. Comput. Appl.*, vol. 141, no. 9, 2016.

[9] N. Boussoualim and Y. Aklouf, "Evaluation and selection of saas product based on user preferences," in *Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2015 Third International Conference on*. IEEE, 2015, pp. 299–308.

[10] M. Naseer and M. Nazar, "A framework for selection of saas by evaluating the quality of freemium model," in *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, Aug 2016, pp. 78–82.

[11] M. A. Ikram, N. Sharma, M. Raza, and F. K. Hussain, "Dynamic ranking system of cloud saas based on consumer preferences - find saas m2nfcp," in *Advanced Information Networking and Applications*, L. Barolli, M. Takizawa, F. Xhafa, and T. Enokido, Eds. Cham: Springer International Publishing, 2019, pp. 1000–1010.

[12] M. A. Ikram and F. K. Hussain, "Software as a service (saas) service selection based on measuring the shortest distance to the consumer's preferences," in *Advances in Internet, Data & Web Technologies, The 6th International Conference on Emerging Internet, Data & Web Technologies, EIDWT-2018, Tirana, Albania, March 15-17, 2018.*, 2018, pp. 403–415. [Online]. Available: https://doi.org/10.1007/978-3-319-75928-9_36

[13] M. I. Fallatah and M. Ikram, "Selecting the right erp system for smes: An intelligent ranking engine of cloud saas service providers based on fuzziness quality attributes," *International Journal of Computer Science & Network Security*, vol. 21, no. 6, pp. 35–46, 2021.

[14] L. Burkon, "Quality of service attributes for software as a service," *Journal of Systems Integration*, vol. 4, no. 3, p. 38, 2013.

[15] G. J. Klir, "Fuzzy arithmetic with requisite constraints," *Fuzzy Sets and Systems*, vol. 91, no. 2, pp. 165 – 175, 1997, fuzzy Arithmetic. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165011497001383

[16] W. Van Leekwijck and E. E. Kerre, "Defuzzification: criteria and classification," *Fuzzy sets and systems*, vol. 108, no. 2, pp. 159–178, 1999.

[17] F. Herrera, M. Lozano, and J. L. Verdegay, "Tuning fuzzy logic controllers by genetic algorithms," *International Journal of Approximate Reasoning*, vol. 12, no. 3-4, pp. 299–315, 1995.

[18] K. P. Yoon and C.-L. Hwang, *Multiple attribute decision making: an introduction*. Sage publications, 1995, vol. 104.

Rectenna Design for Enhanced Node Lifetime in Energy Harvesting WSNs

Prakash K Sonwalkar[1], Vijay Kalmani[2]
Research Scholar, VTU Belagavi[1],
Department of Computer Science and Engineering
Jain College of Engineering, Belagavi - India[1,2]

***Abstract*—In a scenario where every possible solution is investigated for sustainability, Energy Harvesting (EH) stands as an undisputed candidate for enhancing the network lifetime in WSNs where node lifetime decides the network's life. Radio Frequency (RF) energy is abundantly available in the ambience among all the available energy sources. Since both information and power are transmitted in an RF signal, EH is possible in the far-field region. At first, we present a novel 4-element rectangular Patch Antenna Array (PAA) design of EH rectenna. The receiving antenna is designed to pick up the radio signal in the RF range (2.45 GHz) from the free space. Then, the H-shape antenna is modified by introducing a circular slot to enhance the bandwidth. The paper compares the results of the basic parameters of the antenna, such as return loss, input impedance, bandwidth, gain, directivity, and efficiency. As a result, the modified H-shaped antenna (with circular slot) has an increased gain from 8.24 dB to 8.32 dB, with a reduced return loss from -10 dB to -16 dB and enhanced bandwidth from 64.8 MHz to 868 MHz. The high gain, large bandwidth, suitably matched impedance for a minor return loss, and high efficiency of the modified H-shaped patch antenna makes it eligible for energy harvesting application.**

***Keywords*—Antenna design; backscattering; beamfroming; energy harvesting; sequential rule; wireless sensor networks**

I. Introduction

With the proliferation of edge devices and extensive study on deployment, WSNs find their applications ranging from remote applications to body area networks. A typical WSN intends to monitor the environment with the aid of sensor(s), micro-controller(s), transceiver data storage, and energy storage facilities (batteries). The battery acts as an energy source for a node, and its power decides the life of a WSN. Therefore, energy harvesting is perceived as an amicable solution for the bottleneck created by the limited lifetime of the battery. Recently, many researchers have attempted to achieve EH with numerous harvesters and energy resources depending on the applications. While there are many sources for EH, such as solar, wind, thermal, vibrational, temperature, and electromagnetic, RF energy is the most abundantly available, especially in urban environments. The massive adoption of RF-EH can be owed to its ubiquitousness and reliability [1]. A basic block diagram representation of RF harvesting system is shown in Fig. 1.

A typical Rectenna consists of a transceiver antenna, optional Low pass filter, matching network, energy conversion unit and load/storage device. This sensitivity of antennas to RF signal induces an AC signal which is fed to the rectifier. The rectifier comprises diode(s) whose fast switching action is exploited to convert AC signal into DC. A low pass filter

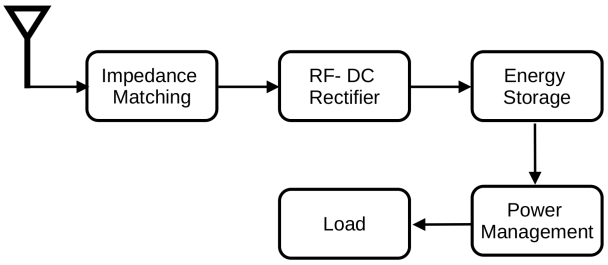


Fig. 1. Typical Block Digram Representation of RF Energy Harvesting System.

is employed for impedance matching between antenna and rectifier to achieve optimal power transfer. For an increased level of output voltages, a voltage multiplier can be employed. The storage and controlling unit provide an uninterrupted power supply. In contrast to other energy harvesters, RF harvesters are robust as they don't require any mechanical movements [2].

RF is an ambient source of energy, arising due to the radiations from TV broadcast, Radio (FM and AM), wireless LAN, Wireless Fidelity (Wi-Fi), and cellular transceiver stations [3]. Although ambient signals can be harvested with simple electronic circuitry, there are many challenges to be addressed by RF harvester:

- Since RF signals are available with wide frequencies, the RF harvester must ensure proper impedance matching for maximum power transfer.
- The RF should employ large broadband antennas to harvest valuable energy from the signals spread over a broad spectrum.
- The harvesting circuits must be positioned close to the RF power source since the ambient levels are deficient.
- The low energy density and low efficiency demand a dedicated RF energy supply as even a high-gain antenna cannot generate enough power densities.

With small-sized, high gain, and impedance-matched broadband antennas along with a reliable system for RF energy supply, energy harvesting in low power WSNs seem to be more promising and feasible.With the stupendous growth in mobile phones and Wi-Fi networks, RF energy has become pervasive and is significant in urban areas [4]. Wireless Power

Transmission (WPT) can be classified into three categories, as described in Fig. 2.

The first category refers to near field inductive or resonant coupling. This reactive phenomenon occurs between two entities where the primary coil transfers power to the secondary. It is suitable for wireless charging of devices separated by few centimetres. The second category refers to far-field directive powering. Here power can be transmitted (in the form of RF) between two entities far from each other but has established a Line-of-Sight (LoS). For example, RF energy can be harvested from mobile phones in proximity, potentially providing power-on-demand for short-range sensing applications. This WPT is employed for the intentional powering of sensors equipped with a rectenna [5]. The third category refers to far-field EH where the receiver doesn't know where the RF energy is emitted (no LoS between the base station and the harvesting device). High gain antennas with wide beamwidth and wide-band resonance are employed for enhanced and efficient energy harvesting in long-range operation.

The selection of the type of rectenna and the entire energy harvesting system varies from application to application [6], [7], [8]. Rectenna is a combination of rectifier and antenna. Diodes are used for rectification, while antenna can be either dipole, planar, or microstrip patch. Many attempts have been made to harvest energy from various RF signals. Among all the frequencies, 2.45 GHz is the favourite. Most of our electronic devices love and live on this radio frequency, as it is in ISM bands (license is not required to operate in this band) [9]. It requires small antennas and can operate over long-range (with LoS). Our objective in this paper is double folded: first, review the attempts made in EH from RF signals and second, to design and develop a high-gain broadband antenna for harvesting over 2.45 GHz signal.

The rest of the paper is ordered below: Section II provides a brief background with theoretical foundations and a literature review of rectennas employed in EH. Section III presents designs of 4-element micro-strip patch antennas for efficient EH. Section IV presents the discussion on designed antennas along with the merits and demerits of each design. Paper concludes in section V.

II. Background and Related Work

Affordable and clean energy is the 7th SDG which aims to cater to the rising demands for energy while reducing the carbon footprint and burden on nature [10]. Energy harvesting seems to be one of the best prospects to realize this goal. EH refers to a process of capturing and storing the energy from sources around us that are free to use. EH, also referred to as Energy Scavenging (ES), makes it possible to overcome the inconvenience of frequent replacement of batteries [11] while being less expensive and eco-friendly. EH stands as a viable solution for endless powering of low power loads such as wireless nodes.

Many attempts have been made to design EH schemes based on the availability of energy sources. Among all, RF-EH is most suitable as the energy source is readily and abundantly available in transmitted energy. Other key benefits are being economically viable, eco-friendly, and having small form factor implementation [12]. RF-EH has the potential

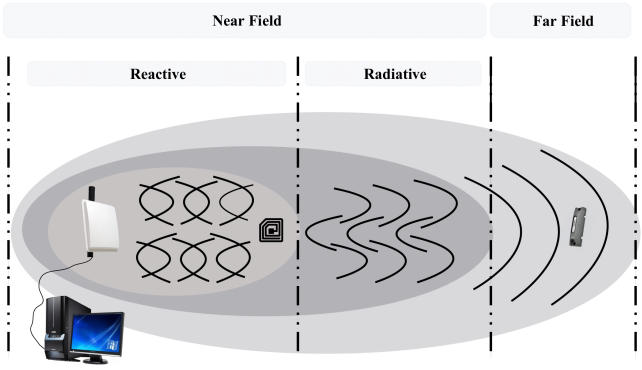


Fig. 2. Three Categories of WPT. Near Field, Far Field - Directional, and Far-Field Ambient Wireless Energy Harvesting

to revolutionize low-power applications - especially WSNs. Excessive use of batteries results in their disposal, causing extreme toxic pollution to the environment [13]. Furthermore, RF-EH can increase the lifetime of nodes and provide power indefinitely [14]. Passive energy scavenging nodes without batteries will be the next generation of WSNs, driven by RF-EH because of its sustainability [15].

A. Theoretical Foundations

A proper understanding of EM waves is must while designing RF-EH system. EM waves largely vary w.r.t. distance, frequency, and conducting environment. Based on the application, designer need to take a call on the parameters of EM waves to make most out of the design. The relation between EM waves and distance from transmitting antenna can be categorized into two segments: near-field and far-field. These two fields are marked by Fraunhofer's distance given by

$$d_f = \frac{2D^2}{\lambda} \tag{1}$$

where d_f is fraunhofer distance, D is the maximum dimension of antenna, and λ is wavelength of EM wave.

For a transceiver antenna, in far field the received power is given by

$$P_R = \frac{P_T G_T G_R \lambda^2}{4\pi R^2} \tag{2}$$

where P_R is power received; $G_T and G_R$ are transmitter and receiver gains respectively.

The RF-DC conversion efficiency is given by

$$\eta = \frac{V_{DC}^2}{P_R R_L} * 100 \tag{3}$$

where V_{DC} is measured DC output voltage, P_R is received RF input power and R_L is resistive load. P_R is given by $P_D * A_{eff}$ where P_D is the RF power density and A_{eff} is effective aperture of antenna [16].

The RF power density for GSM900/1800 MHz is around $0.1\mu W/cm^2$ while for Wi-Fi 2.4 GHz it is around $0.01\mu W/cm^2$ Typically RF power conversion will be around 45% to 50% [17].

B. Energy Harvesting Antenna Design

Design of Energy harvesting antenna has attracted many researchers due to its sustainable and eco-friendly nature. A rectenna is a subsystem of wireless power transfer system which can be designed to function anywhere in the range of 1 GHz to 35 GHz. Many factors such as transmitted power, transmitter gain, received power, receiver gain, conversion efficiency, will dictate the design of an energy harvesting antenna. To enhance the efficiency many other things, need to be considered and implemented such as arrays of antenna and circular polarization of antenna. The resonant frequency of a circular patch antenna is given by

$$f_{r,nm} = \frac{\alpha_{nm} C}{2\pi a_{eff}\sqrt{\varepsilon_{r,eff}}} \tag{4}$$

where α_{nm} is the attenuation, a_{eff} is the effective relative dielectric constant.

Reconfigurable antennas got massive momentum recently due to their tuning, polarization, and selectivity of operating frequency. RF reconfigurability is achieved via dynamic modification of physical structure, thereby attaining polarization diversity. The advantage of automatic frequency tuning to accommodate wideband frequency makes reconfigurable antennas more popular. Though they seem promising, the other constraints like miniaturization, lightweight, beamforming, impedance matching, gain, radiation pattern need to be reworked every time the frequency is switched. While we fine-tune the performances of each module, the combination of all modules should be in harmony and result in efficient Wireless Power Harvesting (WPH) system.

Designing an efficient WPH system involves rigorous testing, making several adjustments, tuning many parameters, and evaluating the entire system. Operating frequency is the prime parameter that dictates the entire design. The operating range also needs to be specified. GHz frequencies are selected for long-range power harvesting, while MHz is sufficient for short-range operations. In a dense environment, electromagnetic waves with very low frequency (in kHz) is preferred. The topology of the electronic circuit(s) such as rectifier and voltage multiplier is decided based on the distance, operating frequency, and the required power output.

Antenna design is an essential part of the entire system design. In addition, rectifier and voltage multiplier design must match the power conversion efficiency. Though the capability of a WPH system critically depends on the antenna, it is one of the overlooked parts of energy harvesting system design. This slight inclination will significantly impact system performance as antennas are selected and deployed irrespective of the operating conditions, material to which it is attached and mobility of the tagged object.

To avoid degraded performance due to improper design of the antenna, the design process has to go through various steps as depicted in fig 3. Understanding the application and deploying environment will enable us to select operating frequency, bandwidth and required antenna parameters. These requirements determine the material for antenna construction

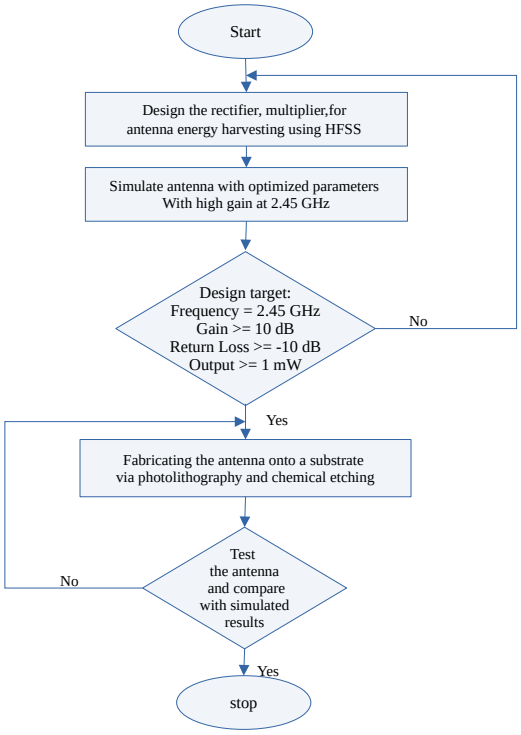


Fig. 3. RF Energy Harvesting Antenna Design Flow

and ASIC packaging. Antenna parametric study and optimization is done until design requirements are met in the simulation. The antenna is first modeled, simulated and optimized on a computer by monitoring the read range, antenna gain, and impedance which provides a good insight into antenna behavior. In the last step of the design process, prototypes are built, and their performance is measured extensively. If design requirement is satisfied, the antenna design is ready. Otherwise, the design is further modified and optimized until conditions are met.

C. Previous Work

Communication antennas have been explored for a century(since World War-I), However, energy harvesting antennas have got the momentum very recently. A narrowband antenna achieves very good energy conversion from RF to DC but can harvest only energy limited to few frequencies. On the other hand, wideband antennas retrieve a large amount of energy but come with large aperture size and poor conversion efficiency. One argument to multi-band antenna is at any given time an antenna cannot be made to resonate at two frequencies [18]. Table I summarizes the prior art of power-harvesting antennas.

Patch antennas have been explored extensively for harvesting energy from RF signal especially at 2.45 GHz [19], [20], [22] and [25]. In [26], antennas with a resonant frequency of 2.45 GHz and 5.8 GHz were designed with Power Conversion Efficiencies (PCEs) of 65% and 46 % @ 10 $\mu W/cm^2$ Two different frequency bands i.e. 900/1800 MHz (for short range) GSM band and 2.4 GHz ISM band were targeted by designing a microstrip antenna with joint feeding line implemented in a Multilayer substrate in [27]. Double patch

TABLE I. COMPARISON OF PUBLISHED WORK REGARDING POWER-HARVESTING ANTENNAS

Ref	Antenna Type	Freq. (GHz)	Gain (dBi)	Dimension (mm)	RF-DC PCE
[19]	Air-substrate Patch	2.45	7	261 * 5	30%
[20]	Patch	2.45	-	100 * 70	73.9% @ 207 $\mu W/cm^2$
[21]	Dual-Linearly Polarized Patch	2.45	7.45 - 7.63	70 * 47.5	78% @ 295.3 $\mu W/cm^2$
[22]	Dual Polarized Patch	2.45	-	100 * 100 * 3.8	82.3% @ 22 dBm
[23]	Dipole	2.45	-	60 * 60 * 60	39% @ 0 dBm
[24]	Microstrip	2.45	8.6	-	83%
[25]	Patch	2.45	4	-	70%
[26]	Patch	2.45	2.19	40 * 43	65% @ 10 $\mu W/cm^2$
[27]	Microstrip	GSM band and ISM (2.45)	-	72 * 94	74% @ 0.3 $\mu W/cm^2$
[28]	Double Patch	1.8 and 2.4	-	40 * 30	19% @ 5 $\mu W/cm^2$
[29]	Single fed Microstrip patch	2.4	7.19	60 * 60	79%
[30]	Microstrip patch Antenna Array 4*4	35	19	-	67% @ 7 mW

*Both refers to inductive or capacitive

antenna was employed by [28] to operate at 1.8 GHz and 2.4 GHz with Simultaneous Wireless Information and Power Transfer (SWIPT) mechanism.

The arrangement of antennas in array is one of the best technique to achieve high gain and to obtain high voltage/current. Another advantage of array antennas over large aperture antennas is that they do not require large breakdown voltage diodes to operate. Connecting antenna array before rectification improves retrieved power at the main beam while placing the array after rectification will expands the ability to retrieve power from wide angles. Combining RF waves before rectification demands for a large breakdown diode, while combining RF waves after rectification, consolidating DC current will be an issue. Series connection of array antennas will enhance voltage whereas parallel fashion is opted for large current. Increasing the number of array elements will yield better outputs but reduces conversion efficiency.

D. Microstrip Patch Antenna Design

1) Estimation of Width of Patch Antenna: Width of microstrip antenna given by

$$W_p = \frac{c}{2f_0\sqrt{\frac{\epsilon_r+1}{2}}} \tag{5}$$

Where f_o is the operating frequency, c is the speed of light in air is 3x10^8 m\sec and ϵ_r is dielectric permitivity of substrate is 4.4.

2) Estimation of Effective Dielectric Constant (ϵ_{eft}): Where h is thickness or height of the substrate which is 1.6mm.

When $\frac{W_p}{h} > 1$,

$$\epsilon_{eft} = \frac{\epsilon_r+1}{2} + \frac{\epsilon_r-1}{2}[(1+12\frac{h}{W_p})^{\frac{-1}{2}} + 0.04(1-\frac{W_p}{h})^2] \tag{6}$$

When $\frac{W_p}{h} < 1$,

$$\epsilon_{eft} = \frac{\epsilon_r+1}{2} + \frac{\epsilon_r-1}{2}[(1+12\frac{h}{W_p})^{\frac{-1}{2}}] \tag{7}$$

3) Estimation of Effective Length (L_{eft}):

$$L_{eft} = \frac{c}{2f_o\sqrt{\epsilon_{eft}}} \tag{8}$$

TABLE II. CALCULATED PARAMETERS FOR THE MICROSTRIP PATCH ANTENNA

Parameters	Value
Effective dielectric constant ϵ_r	2.11
Patch width W	49 mm
Patch length L	39.5 mm
Microstrip Line Length y_0	11.5 mm
Microstrip Line Width W_l	47 mm
Inset gap W_s	47 mm
Width of Substrate W_g	57 mm
Length of Substrate W_g	49 mm

4) Estimation of Length Extension (ΔL):

$$\Delta L = 0.412h\frac{(\epsilon_{eft}+0.3)(\frac{W_p}{h}+0.264)}{(\epsilon_{eft}-0.258)(\frac{W_p}{h}+0.8)} \tag{9}$$

5) Estimation of Actual Length of Proposed Patch Antenna (L_p):

$$L_p = L_{eft} - 2\Delta L \tag{10}$$

6) Estimation of Ground Dimensions (L_g,W_g):

$$L_g = 6h + L_p, W_g = 6h + W_p \tag{11}$$

7) Estimation of Length of the Feed(L_f):

$$L_f = \frac{\lambda_g}{4}, \lambda_g = \frac{\lambda}{\sqrt{\epsilon_{eft}}}, \lambda = \frac{c}{f_o} \tag{12}$$

Where λ_g is Guide wavelength.

III. PROPOSED ENERGY HARVESTING ANTENNA

The basic design of any patch antenna consists of three prominent parts. Ground, Substrate and the Patch. Though only patch works as a functional antenna, Ground and Substrate are added to provide physical support to the patch. In various cases Ground might be the base on which the antenna will be mounted and thus may be eliminated from the design. Substrate, unlike the ground is a must as it affects the radiation and bandwidth of the patch.

By using the above equations, we got the value of each dimension of the antenna, which is showed in Table II. Rectangular Patch antenna array with the dimensions mentioned in Table II is depicted in Fig. 4.

A. 4-Element Rectangular Patch Antenna Array

The geometric parameters are adjusted to observe the variations with respect to the gain, bandwidth, and resonant frequency of the proposed antenna. The patch design incorporates several rectangular slots which are combined to form various shapes. The obtained frequencies is 2.4 GHz. The proposed antenna is designed by using the substrate RT/duroi 5880 substrate form Rogers. This substrate is characterized by 2.20 of dielectric constant with a dissipation factor of around 0.0004. Antenna array is designed using four patch elements with the aim of increasing the gain as shown in Fig. 4. Equidistant placement of the patch elements on the substrate forms a planar array. A feed network is used to connect patch elements and properly designed to enable equal radiation. Among all the available options one side feed network (all patch elements oriented in one direction) provides high gain, low loss, and single major beam with null deviation between electric and magnetic fields.

TABLE III. 4-ELEMENT H-SHAPED MICROSTRIP PATCH ANTENNA ARRAY PROPERTIES

Parameters	Value
Path array dimensions	119.5 * 118 (in mm)
Gain	17.2 dBi
Return Loss	-12.49
Input impedance	$44 + j2.3$
Bandwidth	52 MHz ($\cong$ 2.1%)

B. H-Shaped Patch Antenna Array

Though the antenna is exhibiting acceptable behaviour, the bandwidth is not so superior as compared to previous works. In order to improve the bandwidth one need to tweak around the geometry of antennas without affecting other parameters and properties. Many techniques such as changing or removal (partial) of substrate, and introduction of slots either in radiating patch or ground plane have been investigated. In order to investigate the influence of various shapes and sizes of slots on bandwidth, simulations have been carried out. If we place a slot at the middle of the radiating edge, it may take the form of U or H shape as shown in fig 5 and 6 whose dimensions are mentioned in Tables III and IV. The simulation of H shape antenna array resulted in improved bandwidth.

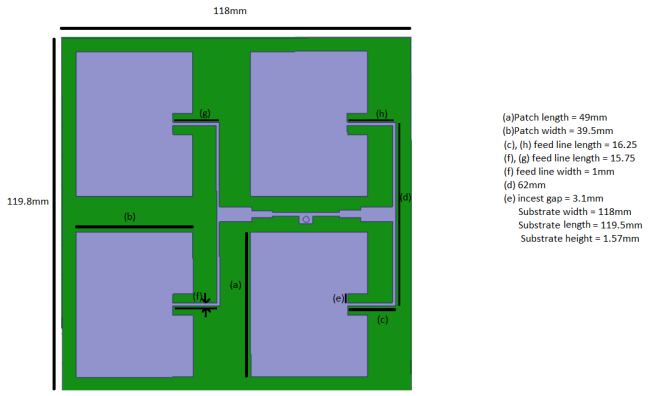


Fig. 4. 4-Element Rectangular Antenna Array

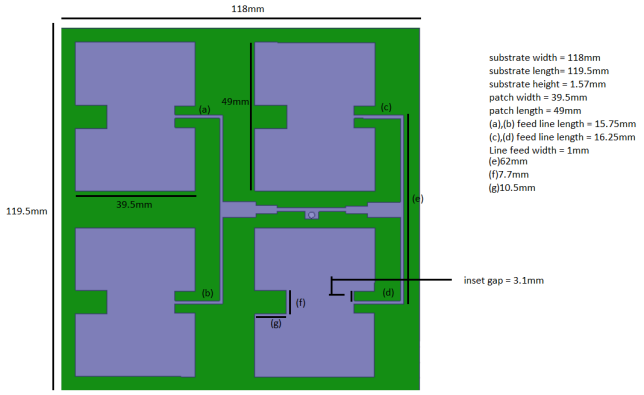


Fig. 5. 4-Element H-Shaped Patch Antenna Array

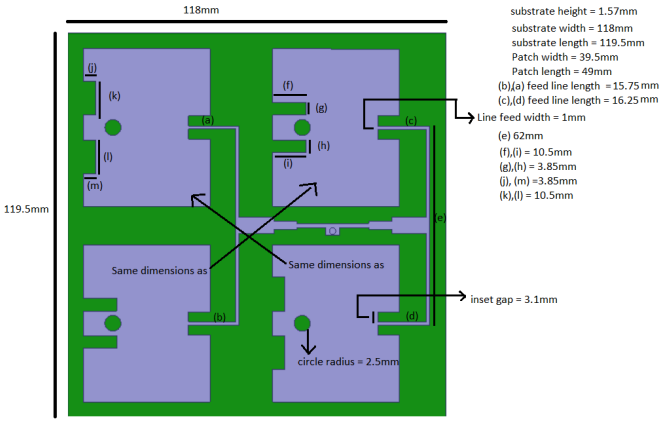


Fig. 6. 4-Element Modified H Shaped Patch Antenna Array

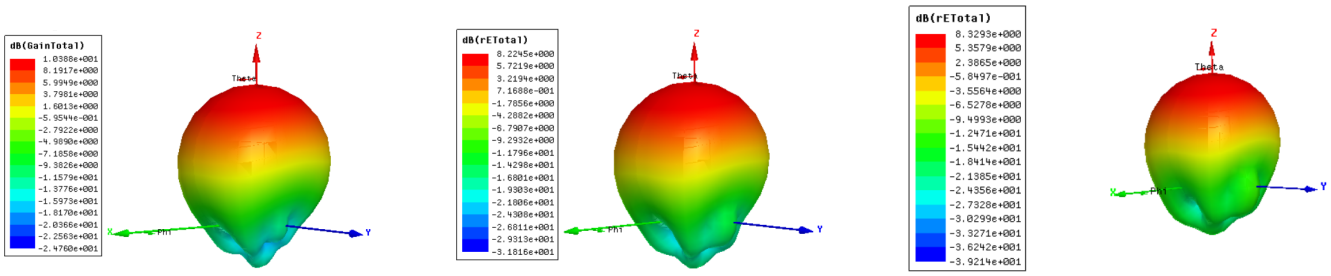
The width is designed as per the equations $W_s = \frac{\lambda}{60}$ and $L_s = \frac{c}{2f\sqrt{\varepsilon_{eff}}} - 2(L + \Delta L - W_s)$.

TABLE IV. 4-ELEMENT MODIFIED H-SHAPED MICROSTRIP PATCH ANTENNA ARRAY PROPERTIES

Parameters	Value
Path array dimensions	119.5 * 118 (in mm)
Slot dimensions	7.7 * 10.5 (in mm)
Gain	17.2 dBi
Return Loss	-12.49
Input impedance	$40 + j5.5$
Bandwidth	64.8 MHz ($\cong$ 2.65%)

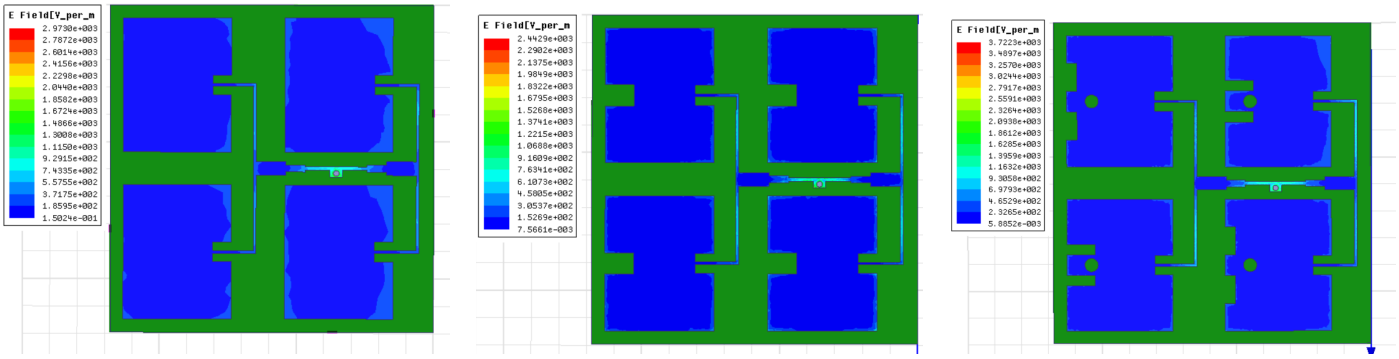
C. Circular Slot, Modified H-Shaped Patch Antenna Array

Though the focus will be on directivity and efficiency while designing an antenna (as shown in Fig. 7 and 8), in the larger picture, the aim will be to have great power reception and conversion (as shown in Fig. 9). Mutiband antennas are designed when energy has to be harvested from RF signals of wide range of frequencies (shown in Fig. 10). The authors major focus on enhancing bandwidth of the rectennas (4-element array) designed to harvest energy at a central frequency of 2.45 GHz.



(a) (b) (c)

Fig. 7. 3D Polar Plots of (a) Single Microstrip Patch Element (b) H Shaped 4-Element Antenna Array (c) Modified H Shaped 4-Element Antenna Array



(a) (b) (c)

Fig. 8. Current Distribution of (a) Single Microstrip Patch Element (b) H Shaped 4-Element Antenna Array (c) Modified H Shaped 4-Element Antenna Array

IV. Results and Discussion

The IoT using wireless motes has perpetuated the demand for self-reliant electronics. Recent research has emphasized fulfilling this requirement via energy conservation. The energy crisis of these remotely placed devices needs to be taken care of. The energy crisis can be studied at various levels, either at energy resources level (choosing an appropriate and abundantly available energy resource), or at energy conservation level (energy-transformation mechanisms), or energy storage level (power management), or at energy consumption level (harvested energy is consumed responsibly).

Energy sources available for harvesting are thermal, solar, vibrational, chemical, Radio frequency, electromagnetic, and mechanical. All the sources demand bulkier devices with mechanical movements for harnessing the energy except RF. In contrast to all available sources, RF is ubiquitous, readily available, and is present in the ambience due to signal transmissions by all wireless transmitters. RF energy harvesting is much suited for IoT devices as there is huge restriction on size of both energy conversion devices and energy storage devices.

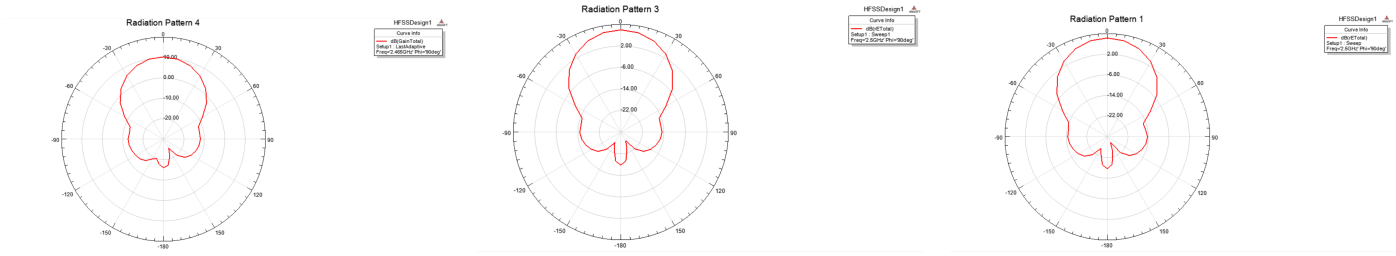
RF energy harvesting has garnered significant attention due to its consistent availability in the ambience from radio, TV, Cellular and Wi-Fi communications. Many energy-saving mechanisms are being investigated, including Radio optimization, Data optimization, scheduling schemes, Routing and Topology Control, and messaging protocols. Among all options, Radio optimization has shown huge potential in energy saving as it deals with energy harvesting. Radio optimization tries to save energy via transmission power control, Directional antenna, and Cognitive Radio.

Wireless transmission of energy has no bounds. Wireless power transfer is the transmission of electrical energy from a transmitter connected to a power source via beamforming to one or more receivers without power cords. At the receiver, the electromagnetic signal is converted back to an electric current and then used by either 1) inductive, capacitive or resonant reactive near-field coupling, or 2) far-field directive power beamforming, or 3) far-field non-directive power transfer.

Since near field and far field with line-of-sign are quiet conducive for energy harvesting, it is implied that much of the research should be focused on the third type i.e. far-field non-LoS WPT. The two challenges in such deployment is the low power densities of incident power and the dynamics of position and orientation of the reciever. The sudden variations in the location brings in much chaos in the power levels which can be addressed by a designing a recitifier capable of operating over a wide-range of incident power. The challenge of low power densities can be partially overcome by having high RF-DC power conversion efficiency (PCE). But it is should be noted that, if more energy is sucked or scavenged from the ambience, the RF-DC will result in much higher PCE. Hence, Rectennas (reciever antennas with rectifiers) need to be designed with broad band to scavenge large amount of low power energy from the ambience such that the voltage multiplier will boost the level and rectifier will take care of the variation in levels.

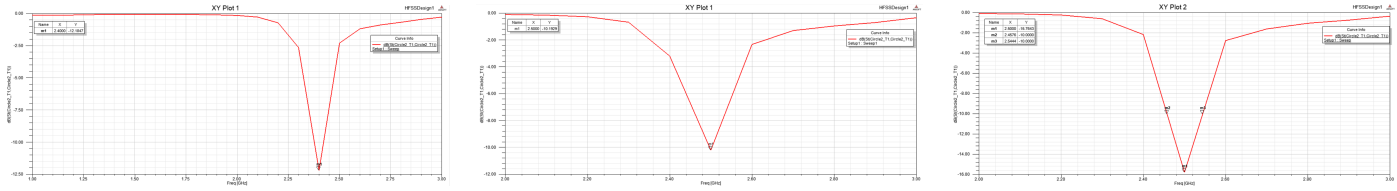
The learning from various sections of this work can be summarized as below:

1) Energy concerns in IoT devices can be rightly addressed by employing energy harvesting mechanisms.
2) Among all available sources of energy, RF energy



(a) (b) (c)

Fig. 9. Radiation Pattern of (a) Single Microstrip Patch Element (b) H Shaped 4-Element Antenna Array (c) Modified H Shaped 4-Element Antenna Array



(a) (b) (c)

Fig. 10. Return Loss of (a) Single Microstrip Patch Element (b) H Shaped 4-Element Antenna Array (c) Modified H Shaped 4-Element Antenna Array

TABLE V. 4-ELEMENT PATCH ANTENNA ARRAY PROPERTIES

Parameters	Value of Rectangular Patch antenna array	Value of H-shaped Patch antenna array	Modified H-shaped antenna with circular slot
Path array dimensions	119.5 * 118 (in mm)	119.5 * 118 (in mm)	119.5 * 118 (in mm)
Slot dimensions	Not applicable	7.7 * 10.5 (in mm)	7.7 * 10.5 (in mm)
Gain	10.4 dBi	8.225 dBi	8.33 dBi
Return Loss	-12.49	-10.49	-15.49
Bandwidth	258 MHz (≃ 10.75%)	64.8 MHz (≃ 2.65%)	868 MHz (≃ 36.17 %)

harvesting is best due to its ubiquitousness, and simple design requirement without any mechanical movements and no demand for large storage.

3) Among all available ambient frequencies, 2.4 GHz is best for long-range wireless power transmission.
4) For a central operating frequency of 2.4 GHz, Microstrip patch antenna is best suited.
5) One side feed network is provides better results.
6) 4-element antenna array is best arrangement for energy harvesting in low-power applications.
7) Creating a circular slot is best option for increasing the bandwidth instead of going for multi-band antennas (which have their own limitation of switching) as demonstrated in this work.

This work has attempted to demonstrate the development of receiving antenna array design starting from a basic rectangular patch antenna. A simple rectangular patch antenna is designed with 4 elements to form an array. The parameters are optimized by the equations defined in Section II, taken from [16]. The centre frequency of 2.4 GHz is accurately achieved. Motivated by this, the authors have attempted to recreate a H-shaped antenna with small improvement in bandwidth from the work [31]. Here a small H shaped antenna, based on tuned slot size is designed. The simulation result shows an incremental change in bandwidth i.e. from 2.1% to 2.65%. Even if we appreciate the delta enhancement, the practical results were not so encouraging as demonstrated in [31]. Therefore in-order to have large bandwidth, the authors have introduced a circular slot in the patch antenna. This inclusion of circular slot has shown remarkable change in bandwidth i.e. from 52.2 MHz (for rectangular patch antenna array) and 64.8 MHz (for H-shaped patch antenna array) to a large bandwidth of 868 MHz. This accounts to 36.17% bandwidth against the 2.65% of H-shaped antenna. This new antenna array can be explored for intelligent beamforming [32]. The comparison of results is tabulated in Table V.

V. CONCLUSION

Given the SDGs, energy crisis is inevitable with the way resources are being exploited. In a low power device placed remotely, energy scavenging is the preferred mechanism to enhance the lifetime of the node. This work considers RF signals to harvest at 2.4 GHz, which is readily available and free to use. The antenna design at this frequency is selected to be microstrip patch with a suitable one side feed network. The work has considered bandwidth to enhance the power conversion efficiency by designing a wideband rectenna with 4-element arrangement for energy harvesting in low-power devices such as IoT devices, Radio Frequency Identifier (RFID) and remote wireless motes.

Energy concerns in IoT devices can be rightly addressed by employing energy harvesting mechanisms. Among all available sources of energy, RF energy harvesting is best due to its ubiquitousness, and simple design requirement without any mechanical movements and no demand for large storage. In all the available ambient frequencies, 2.4 GHz is best for long-range wireless power transmission. For a central operating frequency of 2.4 GHz, Microstrip patch antenna is best suited. One side feed network provides better results. 4-element antenna array is best arrangement for energy harvesting in

low-power applications. Creating a circular slot is best option for increasing the bandwidth instead of going for multi-band antennas (which have their own limitation of switching) as demonstrated in this work. The simple flow from rectangular antenna to the circularly slotted modified H-shape antenna along with the theoretical foundations and the antenna design flow chart, this work acts as a primer for any communication engineer enthusiast to start simulating various slots and enhance various properties of antennas without affecting the other parameters. The authors are confident that the fabricated antenna would give better results and may provide a bandwidth enhancement of at least 20% while considering all non-linearities and implementation losses.

This work has paved a way towards radio optimization and can be extended to transmitter side, where beamforming for energy transmission with receiver location aware precoding can be explored. The authors are confident that the fabricated antenna would give better results and may provide a bandwidth enhancement of at least 20% while considering all non-linearities and implementation losses. The authors also look forward to work on other modules of rectenna: Rectifier, voltage multiplier, power divider, and power management schemes in Wireless Sensor Networks.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of this work, and the Doctoral Committee members of VTU, Belagavi, whose suggestions and insights have shaped this work.

REFERENCES

[1] Adu-Manu, Kofi Sarpong, et al. "Energy-harvesting wireless sensor networks (EH-WSNs) A review." ACM Transactions on Sensor Networks (TOSN) 14.2 (2018): 1-50.

[2] Zeadally, Sherali, et al. "Design architectures for energy harvesting in the Internet of Things." Renewable and Sustainable Energy Reviews 128 (2020): 109901.

[3] Sangare, Fahira, and Zhu Han. "RF Energy Harvesting Networks: Existing Techniques and Hardware Technology." Wireless Information and Power Transfer: A New Paradigm for Green Communications. Springer, Cham, 2018. 189-239.

[4] Khemar, Adel, et al. "Design and experiments of a dual-band rectenna for ambient RF energy harvesting in urban environments." IET Microwaves, Antennas & Propagation 12.1 (2018): 49-55.

[5] Costanzo, Alessandra, and Diego Masotti. "Wirelessly powering: An enabling technology for zero-power sensors, IoT and D2D communication." 2015 IEEE MTT-S International Microwave Symposium. IEEE, 2015.

[6] Sanil, Nischal, Pasumarthy Ankith Naga Venkat, and Mohammed Riyaz Ahmed. "Design and performance analysis of multiband microstrip antennas for IoT applications via satellite communication." 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, 2018.

[7] Punith, S., et al. "A Novel Multiband Microstrip Patch Antenna for 5G Communications." Procedia Computer Science 171 (2020): 2080-2086.

[8] Sanil, Nischal, Pasumarthy Ankith Naga Venkat, and Mohammed Riyaz Ahmed. "Design of an U shaped slotted patch antenna for RFID Vehicle Identification." 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, 2018.

[9] Firestone, Mary. Wireless technology. Lerner Publications, 2008.

[10] Assembly, General. "Sustainable development goals." SDGs Transform Our World 2030 (2015).

[11] Anjum, Shaik Shabana, et al. "Energy management in RFID-sensor networks: Taxonomy and challenges." IEEE Internet of Things Journal 6.1 (2017): 250-266.

[12] Williams, Alexander J., et al. "Survey of Energy Harvesting Technologies for Wireless Sensor Networks." IEEE Access (2021).

[13] Dehghani-Sanij, A. R., et al. "Study of energy storage systems and environmental challenges of batteries." Renewable and Sustainable Energy Reviews 104 (2019): 192-208.

[14] Ahmed, Sheeraz, et al. "Energy harvesting techniques for routing issues in wireless sensor networks." International Journal of Grid and Utility Computing 10.1 (2019): 10-21.

[15] Perera, Tharindu D. Ponnimbaduge, et al. "Simultaneous wireless information and power transfer (SWIPT): Recent advances and future challenges." IEEE Communications Surveys & Tutorials 20.1 (2017): 264-302.

[16] Balanis, Constantine A. Antenna theory: analysis and design. John wiley & sons, 2015.

[17] Piñuela, Manuel, Paul D. Mitcheson, and Stepan Lucyszyn. "Ambient RF energy harvesting in urban and semi-urban environments." IEEE Transactions on microwave theory and techniques 61.7 (2013): 2715-2726.

[18] Kandakatla, Radha Anil, and Mohammed Riyaz Ahmed. "Design and Performance Analysis of Dual-band Microstrip patch antennas for Smart Apparels." 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, 2018.

[19] Momenroodaki, Parisa, Ricardo Dias Fernandes, and Zoya Popović. "Air-substrate compact high gain rectennas for low RF power harvesting." 2016 10th European conference on antennas and propagation (EuCAP). IEEE, 2016.

[20] Sun, Hucheng. "An enhanced rectenna using differentially-fed rectifier for wireless power transmission." IEEE Antennas and Wireless Propagation Letters 15 (2015): 32-35.

[21] Sun, Hucheng, and Wen Geyi. "A new rectenna with all-polarization-receiving capability for wireless power transmission." IEEE Antennas and Wireless Propagation Letters 15 (2015): 814-817.

[22] Chou, Jui-Hung, et al. "All polarization receiving rectenna with harmonic rejection property for wireless power transmission." IEEE Transactions on Antennas and Propagation 62.10 (2014): 5242-5249.

[23] Niotaki, Kyriaki, et al. "A compact dual-band rectenna using slot-loaded dual band folded dipole antenna." IEEE Antennas and Wireless Propagation Letters 12 (2013): 1634-1637.

[24] Sun, Hucheng, et al. "Design of a high-efficiency 2.45-GHz rectenna for low-input-power energy harvesting." IEEE Antennas and Wireless Propagation Letters 11 (2012): 929-932.

[25] Olgun, Ugur, Chi-Chih Chen, and John L. Volakis. "Wireless power harvesting with planar rectennas for 2.45 GHz RFIDs." 2010 URSI International Symposium on Electromagnetic Theory. IEEE, 2010.

[26] Ren, Yu-Jiun, Muhammad F. Farooqui, and Kai Chang. "A compact dual-frequency rectifying antenna with high-orders harmonic-rejection." IEEE Transactions on Antennas and Propagation 55.7 (2007): 2110-2113.

[27] Shuvo, Md Abdul Kader, and Md Mahmudul Hasan. "Multi-Band Microstrip Antenna Design for Wireless Energy Harvesting." American Journal of Energy and Environment 3.1 (2018).

[28] Silva, Vinícius S., et al. "Double Patch Antenna Array for Communication and Out-of-band RF Energy Harvesting." Journal of Microwaves, Optoelectronics and Electromagnetic Applications 19 (2020): 356-365.

[29] Kumar, N. Rajesh, and P. D. Sathya. "Design of RF Energy Harvesting Patch Antenna for Wireless Communications."

[30] Mavaddat, Ali, Seyyed Hossein Mohseni Armaki, and Ali Reza Erfanian. "Millimeter-Wave Energy Harvesting Using 4×4 Microstrip Patch Antenna Array." IEEE Antennas and wireless propagation letters 14 (2014): 515-518.

[31] Chaour, Issam, Ahmed Fakhfakh, and Olfa Kanoun. "Patch Antenna Array for RF Energy Harvesting Systems in 2.4 GHz WLAN Frequency Band." 2018 15th International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, 2018.

[32] Nalband, Abdul Haq, Mrinal Sarvagya, and Mohammed Riyaz Ahmed. "Optimal hybrid precoding for millimeter wave massive MIMO systems." Procedia Computer Science 171 (2020): 810-819.

Politicians-based Deep Learning Models for Detecting News, Authors and Media Political Ideology

Khudran M. Alzhrani
Department of Information Systems
Al-Qunfudhah Computing College, Umm Al-Qura University
Al-Qunfudhah, Mecca, Saudi Arabia

***Abstract*—Non-partisanship is one of the qualities that contribute to journalistic objectivity. Factual reporting alone cannot combat political polarization in the news media. News framing, agenda settings, and priming are influence mechanisms that lead to political polarization, but they are hard to identify. This paper attempts to automate the detection of two political science concepts in news coverage: politician personalization and political ideology. Politicians' news coverage personalization is a concept that encompasses one more of the influence mechanisms. Political ideologies are often associated with controversial topics such as abortion and health insurance. However, the paper prove that politicians' personalization is related to the political ideology of the news articles. Constructing deep neural network models based on politicians' personalization improved the performance of political ideology detection models. Also, deep networks models could predict news articles' politician personalization with a high F1 score. Despite being trained on less data, personalized-based deep networks proved to be more capable of capturing the ideology of news articles than other non-personalized models. The dataset consists of two politician personalization labels, namely Obama and Trump, and two political ideology labels, Democrat and Republican. The results showed that politicians' personalization and political polarization exist in news articles, authors, and media sources.**

***Keywords*—Deep neural networks; text classification; political ideology; politician personalization**

I. Introduction

This paper examines the relationship between two political science concepts that usually addressed separately, news articles' political ideology and personalization. Political ideology of an article can be predicted based on its content, publisher ideology, or authors leanings. Our main hypothesis is that by constructing intelligent models trained on politician centered articles will improve their performance in detecting articles' political ideologies. Personalization attributes such as politician visibility, personal traits, and reoccurring topics form a pattern that statistical models learn to distinguish personalized articles. articles' tags are used in building the dataset, which indicate that the politician is directly related to the topic covered in the article. In the Presidential Dataset (Section III), Obama was mentioned in 23.6% of the articles that were tagged with Trump. Despite that, our models were able to detect articles personalization and the ideology of personalized articles effectively. Nonetheless, introducing intelligent models to detect articles' ideology based on their personalization proved to advance their performance.

One would question the motivation of paper by pointing out that the readers could identify politicians mentioned in the article and be aware of the article's writer or publisher's political alignment at the same time. However, not all readers are attentive [1], and political leanings of news sources or writers are not always known. Also, detailed articles might have few mentions of a politician and not be personalized. Some readers do not go beyond initial information exposure of shared articles, or cropped news on a website and news feeds [2]. Also, websites' structure differs from one another, meaning not all news websites use web tags. Some websites' main page displays articles in their entirety; others show the articles' title and maybe a snippet of its first paragraph. News agencies share links to the news articles on social media networks. Readers might arrive at conclusions solely based on the articles' headlines or posts on character-limited social networks. Automated personalization detection increases the amount of information available to readers, even without reading the article. The pre-trained models can easily detect who is personalized in this article and notify the reader. Models can be trained on multi-labeled data to detect more than one politician's personalization.

The research's interdisciplinary nature contributes towards bringing artificial intelligence to political science, journalism, and communication. The proposed concept has practical applications for regular readers, news outlets, social networking sites, and politicians. The intelligent models will provide awareness to the reader about personalized articles' ideologies. Labeling the article based on its political ideology helps the reader identify which standpoint the article takes. Hence, the reader is encouraged to seek the other viewpoints to determine which one is more convincing to align themselves to it or take a specific position. For instance, if most of the articles exhibited to a reader were from one political side, the reader's awareness of being in a bubble would increase.

News Recommender Systems (NRS) aims to customize the news articles displayed to the readers based on criteria or techniques that are believed to capture the readers' interest. Researchers have proposed several expert systems to recommend the news from the web [3], [4], news agencies' tweets [5], and heterogeneous data tailored for journalists [6]. Moreover, social networks integrate news recommendation systems into their platform, such as Facebook news feed, to suggest news articles based on users' behavior, etc. [7]. Due to deep neural networks' ability to overcome some of the

limitations in the traditional techniques, the incorporation of deep neural networks into news recommendation systems have been more frequent in recent years [8], [9], [10]. Regardless of the adopted methods to build the news recommendation systems, researchers pointed out that such systems increase pro-attitudinal selective exposure of political news [11], [12], [13]. Articles' personalization and ideology detection models can help the news recommendation systems mitigate pro-attitudinal selective exposure.

Deep transfer learning for texts enables knowledge sharing between a source domain with a sufficient amount of labeled data to target domains that suffer from labeled data insufficiency. The transfer could be achieved by different means, e.g., pre-trained word embedding [14], instance-based [15], and adversarial-based [16]. Because the Presidential Datasets are reconstructed to target articles, authors, and sources, transferring the DNN to multiple domains is possible. Personalization of politicians is transferable to other personalization types as CEOs' news media coverage [17]. Researchers pointed out that the media treats CEOs as brands and that preserving the CEO's image improves products and firms' value [18]. Personality prediction [19] and framing and agenda-setting detection [20] are other examples of transfer learning destinations. Other domains that are suitable for transfer learning from ideology detection models are emotion detection [21], top-specific opinion classification [22], and news bias identification [23].

In this work, a new approach is proposed to improve the performance of news political ideology detection models by building models with feature space extracted from politicians' personalized articles. Politicians' personalized-based models for political ideology detection were able to achieve higher or match the performance of other non-personalized political detection models, which trained on much more data. In addition, The experiments proved that it is possible to detect articles politicians' personalization automatically. The detection models were evaluated against news articles, authors, and media sources to examine the relations between the two political concepts. Finally, a statistical analysis of the presidential dataset from the political personalization perspective is provided.

The reminaing of this paper is structured as follows: Section II briefly reviews news personalization literature and describes how this paper's contribution can fit in this research field. We analyze the statistics related to articles' personalization in the Presidential data and its relation to political ideology in Section III. Description of research models and experimental setup provided in Section IV. Section V report and discuss results of articles' personalization and ideology detection. Finally, the paper is concluded in Section VI.

II. Related Work

The related work of the automated texts' ideology detection is extensively reviewed in [24]; therefore, it will not be covered in this paper because. In this section, the literature of political personalization in the news media is reviewed.

Personalization in the literature comes in many shapes and forms. Personalization could mean delivering services [25], advertisements [26], or educational content [27] that best suited for a targeted individual. In law, personalization is substituting a uniform law for one tailored to an individual's preference, characteristics, or circumstances [28]. Other fields have several definitions for personalization as in marketing [29] and e-commerce [30], leading to confusion or ambiguity. Similarly, scholars in political science and communication view personalization from different perspectives. Researchers have examined the existence of personalization in various types of political institutions [31]. Others focused on studying personalization impact on the leaders' [32] or the public [33] political behavior. Numerous studies investigated the role of personalization in political campaigns and elections [34], [33]. Some of the political scientists debated that the rise of personalization is beneficial for modern democracies—the following are some of the observations that support this claim. Leaders will advocate the party's message and increase citizens' political engagement through direct online communication [35]. Voters' attachment to parties weaken due to increased political personalization; hence, voters would be more willing to vote for a different party in subsequent elections [36]. On the other hand, some will argue that personalization brings more harm than good to democracy [37]. For example, loosen the ties between voters and parties might move voters' and political parties' attention from local to national elections [36]. The literature on political personalization is extensive and diverse. This paper is closely related to papers that address presidential personalization in the online news media.

Peoples' political disagreements on the personal level manifest itself by voting for candidates who align with their political ideology as in representative democracy. The political parties mostly nominate their preferred candidates according to their set of criteria. These candidates roles revolves on advocating and implementing the party's program once they get in the government. Therefore, to some degree, the voter, candidate, and the party all have the same political ideology; hence, voters tend to stick with one party to help advance its policies. Mughan [38] stated that in the sixties, the United Kingdom and United States voters' interest shifted from the political parties to an individual politician's personalities, especially those in the high government positions. The media attention moves from the party loyalty to the particular politician's personalities at the top to lead, namely presidentialization [38], [39], [40].

The use of presidentialization as terminology to describe the UK's prime ministries nominees' influence over the electoral process and the elected prime minister rise in power over their cabinet members is debated [41]. Other political scientists prefer the term prime ministerialisation [42] or personalization [43]. Furthermore, individuals attribute populism mounting to personalized politics, where political leaders with compelling personalities appeal to a broader range of voters [44], [45]. Although the degree of political personalization differs depending on multiple variables such as the number of competing parties [46], it is common to find political personalization across western democracies [47], [48], [49], [36]. In her comprehensive book [47], Bittner classified the divergent research of political personalization into five categories: leaders' selection, leaders' traits, leaders' evaluation, leaders' impact on electoral outcomes, and information sources. The media is considered one of the primary information sources available to voters, and hence the criticality of the media's depiction of a political candidate is critical.

The waning of the power of monolithic politics gave rise to the media coverage of individual politicians. Aelst et al. [50] developed a model to organize the political news personalization studies into two dimensions logically. The first dimension compares political news coverage of individual politicians and political institutions, leading to a dimension known as individualization. The second dimension emphasizes the change in politicians' media coverage as a government official to a private citizen; hence, this dimension is labeled privatization.

The individualization stems from two decentralized branches, which examine any or all politicians' media visibility regardless of their position [50], [51]. It also includes centralized, which study's the media's visibility regarding different political leaders and their characteristics [52], [53], [46]. Relativity different take on the distinctions between centralized and decentralized media personalization articulated in [54] defined centralized personalization as the media visibility of an upward shift towards politicians at the top, On the other hand, researchers described decentralized personalization as a downward shift towards politicians in lower positions or parties. Privatization consists of two sub-dimensions [50], [55], personal characteristics and personal life. Personal characteristics refer to increased media coverage of a politician's traits rather than his political elements. Personal characteristics make people aware of a politician's positive or negative aspects as those who focus on those aspects aim at attaining certain political goals. On the other hand, personal life, which addresses media coverage, shifts to politician personal activities and interest. Research in politicians' traits is inconclusive since the dissimilarities between personal and political traits are clouded [56].

Others argue that personal disclosures in the media could be politicized [50]. Now that personalization concepts laid out in the previous paragraphs, one cannot place personalized media detection into these dimensions and subdimensions. We are not aware of any paper that researched the problem of news media personalization detection. However, one can link this work to some of those subdimensions. In personalized media detection, the deep network models learn politicians' personalities and political traits with other attributes to form a pattern in which one quickly identifies personalized articles. This research paper also touches on presidentialization or centralized personalization by experimenting with personalized articles regarding two US presidents,, namely President Donald and Trump and Barrack Obama. The two presidents have massive information attributed to them in the media and different other websites. The information is mostly published to sensitize the public about their aspects or public aspects crucial in politics. Despite that, personalized media detection should be distinguished from other previously known personalization dimensions. However, this distinguishing aspect does not eliminate the possibility of incorporating personalized media detection in different types of news personalization.

This paper is the first research paper that considered articles' political personalization as a dimension to detect ideologized text to the best of our knowledge. On top of that the ideology detection models' performance are evaluated in association with articles' authors and sources.

TABLE I. THE TABLE SHOWS SIZE OF POLITICAL IDEOLOGY AND POLITICAL PERSONALIZATION LABELS IN THE PRESIDENTIAL DATASET.

		Ideology		Personalization	
Set	# Articles	Conservative	Liberal	Obama	Trump
Train	125051	35035 (28%)	90016 (72%)	72256 (58%)	52795 (42%)
Test	53521	15017 (28%)	38504 (72%)	30896 (58%)	22625 (42%)

TABLE II. THE TABLE SHOWS THE TOTAL NUMBER OF TRAINING AND TESTING ARTICLES FOR THE IDEOLOGY CLASSES IN TRUMP AND OBAMA DATA.

	Trump Dataset		Obama Dataset	
Class/Dataset	Train	Test	Train	Test
Conservative	22933 (43.4%)	9712 (43%)	12102 (16.7%)	5305 (17.2%)
Liberal	29862 (56.6%)	12913 (57%)	60154 (83.3%)	25591 (82.8)
Total	52795	22625	72256	30896

III. PERSONALIZATION IN THE PRESIDENTIAL DATASET

The articles in the Presidential dataset [57] were collected from multiple news sources aligned with extreme left and right on the political spectrum. A detailed description of the dataset and the methods used to collect the articles are provided in [24] and listed in Table I. The dataset is a collection of articles written about two U.S. presidents, Trump, and Obama. The two represent two different political ideologies and covered by all news agencies. The paper will focus on personalized articles in the dataset and its relation to the article's ideology, which were not explained in the previous publications. And examine how the dataset is balanced in terms of media attention to the politician and the ideology of media sources that published the articles.

Data imbalance is a common problem in text classification [58], and it hurts the performance of the detection models. As seen in Table (), the presidential dataset can be organized based on politician personalization, and political ideology. From personalization perspective Trump and Obama's articles represent about 42% and 58% of the entire corpus, respectively. On the other hand, 28% and 82% of the articles are labeled with Conservative and Liberal political ideologies, respectively. In comparison, personalized articles are more balanced than ideology articles. The proposed hypothesis assumes that constructing intelligent models trained on political personalized articles to detect political ideologies would improve the detection models performance. Knowing that personalized articles are more balanced than the ideology ones might contribute to this matter. However, one cannot be certain that the same could apply to the size of media coverage of other politicians. Politician serving times, position and other factors can generate more media attention.

Table II lists detailed statistics of personalized corpora. More than 43% of the articles written about Trump are conservative, and only 17% are labeled as conservative in the Obama corpus. As can be seen in Fig. 1, articles labeled as Liberal are more significant than Conservative articles in the Obama and Trump datasets. Moreover, the gap between the size of Conservative and Liberal articles is much more evident for Obama's articles. Therefore, one can assume that intelligent models would be able to detect the ideology of

TABLE III. THE TABLE DISPLAYS NEWS SOURCES, IDEOLOGY ALIGNMENTS, AND THE NUMBER OF ARTICLES IN TRUMP AND OBAMA DATASETS.

News Website	Ideology	Trump Dataset	Obama Dataset
DailyWire [7.57%]	Conservative	11031 [14.62%] (Tr:7797 , Te:3234)	2488 [2.41%] (Tr:1757 , Te:731)
ILoveMyFreedom [4.22%]	Conservative	6350 [8.42%] (Tr:4488 , Te:1862)	1180 [1.14%] (Tr:826 , Te:354)
DailyKos [68%]	Liberal	39822 [52.8%] (Tr:27808 , Te:12014)	81691 [79.19%] (Tr:57327 , Te:24364)
National Review [11.2%]	Conservative	9981 [13.23%] (Tr:6929 , Te:3052)	9945 [9.64%] (Tr:6880 , Te:3065)
TheBlaze [0.62%]	Conservative	614 [0.81%] (Tr:432, Te:182)	494 [0.48%] (Tr:351 , Te:143)
WorldSocialist [3.92%]	Liberal	2953 [3.91%] (Tr:2054 , Te:899)	4054 [3.93%] (Tr:2827 , Te:1227)
NewsBusters [4.4%]	Conservative	4669 [6.19%] (Tr:3287 , Te:1382)	3300 [3.2%] (Tr:2288, Te:1012)
Total	Combined	75420	103152

articles covering Trump better than Obama.

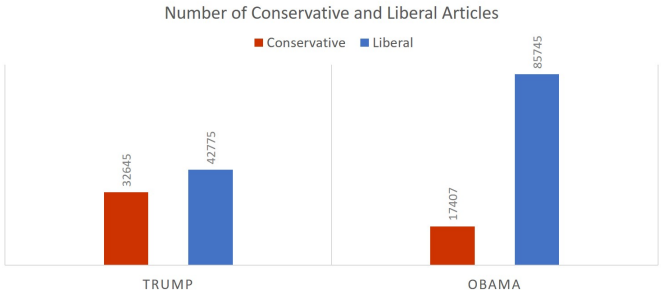


Fig. 1. Personalization V.S. Ideology. The Figure Shows the Number of Conservative and Liberal Articles Tagged with Trump and Obama.

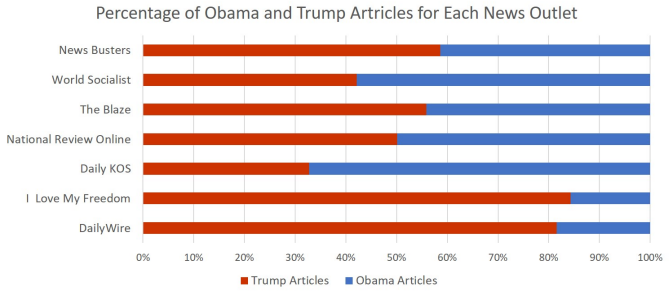


Fig. 2. Percentage of Personalized Articles in Media Sources. Display the Percentage of Articles Tagged with Trump and Obama in each Data Source.

The reason behind the difference in size among personalized and ideology articles can be understood better by exploring the number of articles generated by each media source. Table III presents all the news media sources' political ideology and the number of articles written about Trump and Obama. A large portion of the Presidential dataset comprises of articles from the DailyKos, 68% to be exact.

The difference in coverage becomes apparent in the horizontal bar chart in Fig. 2. DailyKos, a liberal media outlet, represents 52.8 % and 79.19% of Trump and Obama datasets, respectively. The second and last liberal data source, WorldSocialist, is the only media source with consistent representation in the Presidential, Trump, and Obama datasets, which is, on average, 3.92%. The remaining data sources are conservative and mostly have more written articles about Trump than Obama. One can conclude from this plot that all conservative sources, except for National Review Online, have more articles about Trump than Obama. On the other hand, Obama received more attention in the liberal media sources.

Parts of some articles in the Presidential Dataset are in Table IV. These articles' snippets provide the information needed to determine the authors' standpoints on controversial issues such as Obamacare, Iran's nuclear deal, and climate change. For instance, the far-left news source DailyKos article covering an event that took place in the Senate to repeal Obamacare pointed out the failure of the Republicans in the Senate to take health coverage from 16 million people. Another article reporting on the same event but from the conservative point of view stated that the modest Obamacare repeal offered by the Senate Republicans was 'killed' by the Democrats. Other examples are seen in the table and throughout the dataset. The experiment section will discuss the detection models trained on Trump and Obama datasets independently to detect their articles' political ideologies. Also, separate sets of the detection models will be trained on the presidential training set to identify Trump and Obama's personalized articles.

IV. RESEARCH MODELS AND EXPERIMENTAL SETUP

Similar to [24], the following four deep neural network models are employed to detect news' articles politician personalization and political ideology.

1) FastText [59] models enable tasks to word representation fine tuning through averaging word vectors and updating embeddings in the training phase through back-propagation. The word embeddings are then fed to a fully-connected layer with Softmax activation to map articles representation to category labels. Articles word order is ignored during the construction of the text representation, which increase FastText classifiers speed and still achieve relatively good results.
2) Convolutional Neural Networks for the text classification (TextCNN) [60] use multiple convolution layers with different kernel sizes. In text classification, the convolution layers are effective in extracting features over multiple sliding windows from one dimensional inputs. The output of the convolution layers are max-pooled over the entire sequence to identify the most useful features and generate fixed-length vector. The features maps constructed from the max-pooling

TABLE IV. THE TABLE LISTS A SAMPLE OF ARTICLES THAT BELONGS TO THREE CONTROVERSIAL ISSUES IN US POLITICS OBAMACARE, IRAN'S NUCLEAR DEAL, AND CLIMATE CHANGE.

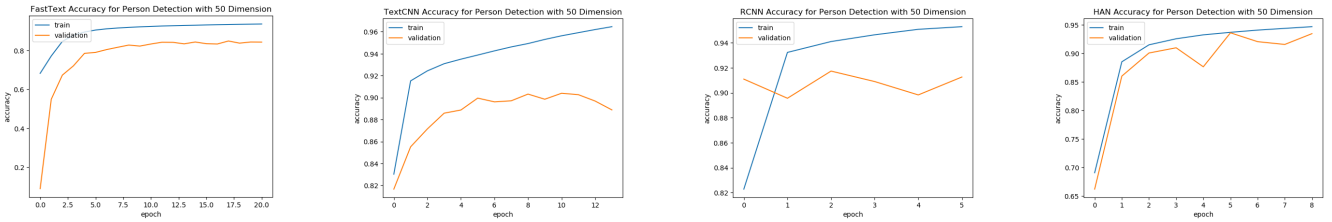
Theme	Personalized/Ideology	Snippet From The Article	Date	Source
Obamacare	Trump/Liberal	Senate Republicans failed in their latest attempt to take health coverage from 16 million people with so-called "skinny repeal".	07-28-2017	DailyKos
	Trump/Conservative	Around midnight Thursday, the Senate Republicans' attempt to pass even the most modest of their Obamacare repeal efforts crashed and burned when "Maverick" John McCain joined the so-called moderate Republicans Susan Collins (ME) and Lisa Murkowski (AK) to vote with the 48 Democrats to kill the "skinny repeal" bill.	07-28-2017	DailyWire
	Obama/Conservative	Democrats from Sen. Chuck Schumer (D-NY) to former President Barack Obama announced this weekend that their first priority after the holidays will be to preserve the Affordable Care Act which was declared unconstitutional (again) Friday by a federal district court judge.	12-16-2018	DailyWire
	Obama/Liberal	President Obama visited Texas last week. He had a chance to visit with Texas families personally impacted by the health care law, Obamacare.	04-18-2014	DailyKos
Iran's Deal	Trump/Liberal	Donald Trump is looking for excuses to pull out of the Iran nuclear deal—after all, it was negotiated under former President Obama, whose accomplishments Trump is looking to wipe out one by one—even though his own top advisers like Defense Secretary James Mattis and Joint Chiefs of Staff Chair Gen. Joseph Dunford have said that the deal is working.	10-10-2017	DailyKos
	Trump/Conservative	Donald Trump was exactly right when he called the Iran deal a "horrible" and "disastrous" agreement.	11-14-2016	NationalReview
	Obama/Conservative	On Sunday's Face the Nation, Washington Post columnist Michael Gerson blasted President Obama for aligning Republicans in Congress with the leadership with Iran who chant "Death to America" simply for opposing the nuclear deal.	08-09-2015	NewsBusters
	Obama/Liberal	A comprehensive deal on Iran's nuclear program has been done, diplomats in Vienna said, bringing to an end a 12-year standoff that had threatened to trigger a new war in the Middle East.	07-14-2015	DailyKos
Climate Change	Trump/Liberal	Much has been made of the fact that Trump's withdrawal from the Paris Climate Agreement makes the United States an international pariah and a business disaster, that it will undermine national security, that his excuses were a series of flat-out lies, and that his real reason was the typical pettiness of his tiny, fragile ego, but the degree to which Trump is Making America Worst cannot be overstated.	06-02-2017	DailyKos
	Trump/Conservative	By withdrawing from the agreement, Trump could restore the Senate's constitutional power to advise on and consent to international treaties.	05-09-2017	NationalReview
	Obama/Conservative	Wednesday's edition of the CBS Evening News chose to re-air portions of chief medical correspondent Dr. Jon LaPook's interview with President Obama on climate change supposedly threatening public health and included LaPook fretting at the end to anchor Scott Pelley that "climate change legislation has stalled in Congress."	04-08-2015	NewsBusters
	Obama/Liberal	Because it worked so well to scuttle the global agreement to prevent Iran from getting a nuclear weapon (not), Republicans are trying the same techniques to undermine the Paris talks to combat climate change.	09-13-2015	DailyKos

layers are stacked in the concatenation layer. The final layer is fully-connected, then a Softmax activation applied on the final layer output resulting in a class prediction. One drawback from this approach is losing the sequential order of texts, and not being able to model more sophisticated structures.

3) Recurrent Convolutional Neural Networks (RCNN) [61] combines recurrent and convolution layers to take the advantages and mitigate the limitations of both layers. RNN captures contextual semantics by constructing local feature maps of text sentences. Three text representation, left context, right context, and standard word embedding are shared during training update with separate outputs. Forward RNN constructs the left side context, and the right side context is generated by a Reverse RNN. All three text representations are merged by a concatenation layer that fed to a convolution layer. Then a max-pooling layer extracts global most influential feature vectors. The final layer is a fully-connected that passes its output through a Softmax activation function for class prediction.
4) Hierarchical Attention Networks (HAN) [62] best suited for document classification due to its two levels attention mechanisms. Fixed-length of input words encoded by Long-short-term-memory (LSTM) layer. The LSTM itself is wrapped in Bidirectional RNN layer to perform backward RNN computation. The following layer is an attention that forms sentences from the most useful words. Time distributed layer wraps the bidirectional encoder and the attention layer. Another bidirectional RNN layer further encodes the processed sequences to construct a document from the most informative sentences; followed by an attention layer. Finally, a fully-connected layer passes its output to a Softmax activation function to compute the probability of document belonging to a class.

The same networks settings, learning configuration, and text prepossessing techniques depicted in [24] are used in this paper for fair comparison with earlier experiments. Prior to the learning phase, the texts are tokenized based on white space as delimiters. All letter cases are lowered, and punctuation removed from the tokenized texts. The remaining texts are sequenced and padded to 400 sequence length. The highest possible number of features is 35000 unigrams. Word embedding size is 50 with random initialization.

As for the neural network settings, Adaptive Moment Estimation (Adam) is the learning optimizer. The optimizer learning rate is set to 0.0001 and 0.9, 0.999 for the optimizer beta_1 and beta_2 decay parameters, respectively. Both politician penalization and political Ideology detection are single classification problem; therefore, Binary Cross-entropy is selected as a loss function. The model that achieved the best accuracy results on the validation set during training process



(a) FastText Personalization Detection Accuracy

(b) TextCNN Personalization Detection Accuracy

(c) RCNN Personalization Detection Accuracy

(d) HAN Personalization Detection Accuracy

Fig. 3. Training and Validation Accuracy of Articles' Politicians Personalization Detection Models

is chosen to evaluate the model performance on the testing set. Accuracy as a metric is only used on validation set, other metrics discussed below evaluated models' performance on the testing set. 50 is maximum number of epochs with 32 batch size. Of the training set 15% is set aside as a validation set to validate model performance posterior to each epoch. With an early-stop scheduler the training will terminate if the validation accuracy did not improve for three consecutive epochs. The networks were built in the Keras platform with TensorFlow as the backend in all the experiments.

During the entire experiment, several metrics evaluated the performance of the neural network models. The best fitted model on the validation set is chosen based on the Accuracy metric, see Equation 1. In binary classification, the accuracy of the model is calculated by dividing the number of correctly predicted examples, which equals to the sum of True Positive (TP) and True Negative, over total number of examples represented by the sum of TP, TN, False Positive (FP), and False Negative (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

While the accuracy metric is a good measure of models performance, it is limited when it comes to imbalanced dataset. Therefore, Precision (Eqn. 2, Recall (Eqn. 3), and F1-score (Eqn. 4) are the measurements metrics used to report the results of detection models on the testing set. Precision computes the fraction of correctly predicted examples of a class among all the examples labeled by the model as the relevant class.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

On the other hand, Recall derive the fraction of correctly predicted examples of a class among all the examples that actually belong to the relevant class.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The harmonic mean of the Precision and Recall or the metric known as F1-score is a metric that overcomes some of the limitations found in other metrics. The F1-score is suitable or imbalanced data and it gives equal importance to Precision and Recall.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

V. Results and Discussion

Constructing ideology detection models trained on personalized set boosts or at least maintain ideology detection performance despite downsizing the training set size. Therefore, this work examines the deep network models' ability to detect articles' personalization. Attaining reasonably satisfactory results in distinguishing among articles centered around particular politician will further prove the existence of political personalization in news media coverage. The deep neural networks trained on articles tagged with politicians, namely Trump and Obama. This paper compares between the results obtained by ideology detection models reported in [24] and personalization detection models. In the subsection V-B, we test our hypothesis and see if personalized detection models are better at detected the ideology articles than non-personalized ones.

A. Politicians' Personalization Detection Results

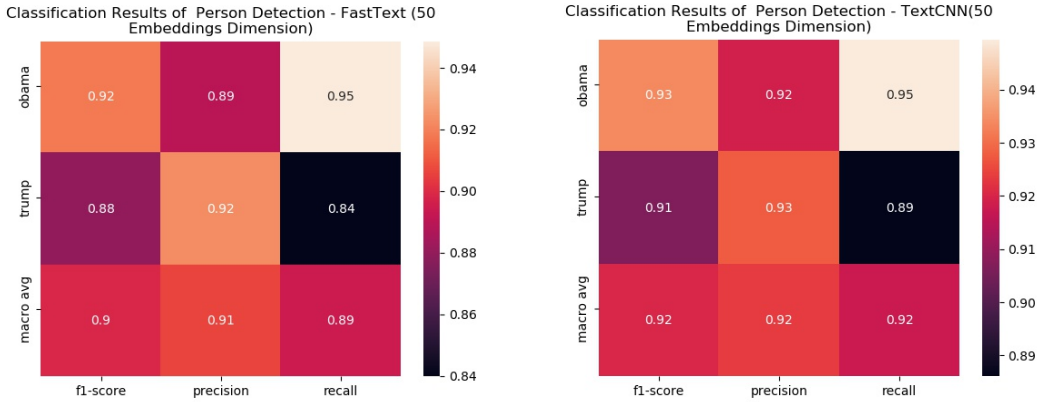
1) Detection Models' Validation Accuracy: It is beneficial to examine the accuracy results of the validation set obtained from the training procedure. The same deep networks employed in the detection of articles' ideology is used, which will assist in observing network behavior trained on the same data but different classes. As depicted in Fig. 3, FastText, Fig. 3a, required 20 epochs before reaching the termination point, which is more than TextCNN 13 epochs, RCNN 5 epochs, and HAN 8 epochs, Fig. 3b, Fig. 3c, and Fig. 3d, respectively.

FastText is the fastest network among the four deep network models. The best accuracy of 93% on the validation set was achieved by HAN, which is a little bit shy from the training accuracy at its peak. On the other hand, TextCNN accuracy decreased in the last three epochs from 90% to around 88%. Finally, RCNN was able to recover from descent in epoch 4 to 92% at the fifth and final epoch. The accuracy results from the validation set do not necessarily reflect the model's performance on the testing set since the models' assessment is done on 15% of random non-stratified training data.

2) Politicians' Personalization in News Articles: This subsection compares the results of the four deep network models based on three metrics, Precision, Recall, and F1-Score. In Fig. 4, heatmap and bar charts display numerical and graphical

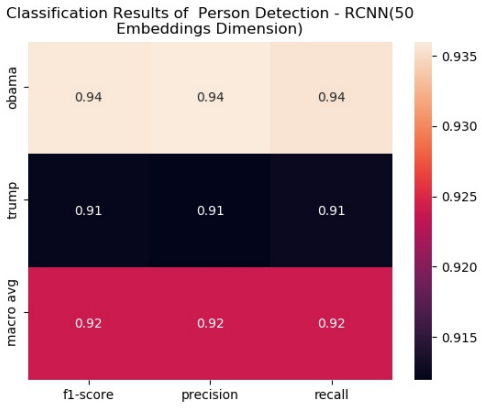
measurement values for detailed comparison. FastText, Fig. 4a, was the least accurate with 0.92 F1-Score for Obama, and 0.88 for Trump personalized articles.

All deep network models are better at detecting articles' personalization than ideology. TextCNN and HAN reported the same F1-Score for both classes, Fig. 4b and 4d, but they differ in Recall and Precision. TextCNN is better at detecting Obama articles with 0.95 Recall and 0.92 Precision, yet HAN is less likely to misclassify articles written about Obama as Trump with 0.91 Recall and 0.95 Precision. However, RCNN, Fig. 4c, outperformed other models at detecting Obama with 0.94 Recall, Precision, and F1-Score. As for Trump's articles, all personalization detection networks got 0.91 F1-Score, 0.93 was the best Recall value recorded by HAN, and 0.93 was the best Precision by TextCNN. In any case, Obama dataset size might be one of the reasons why it received better prediction

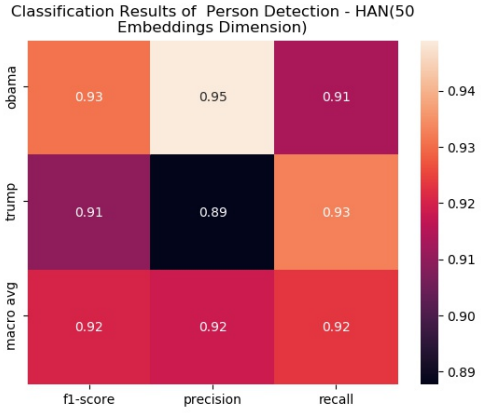


(a) FastText Personalization Detection

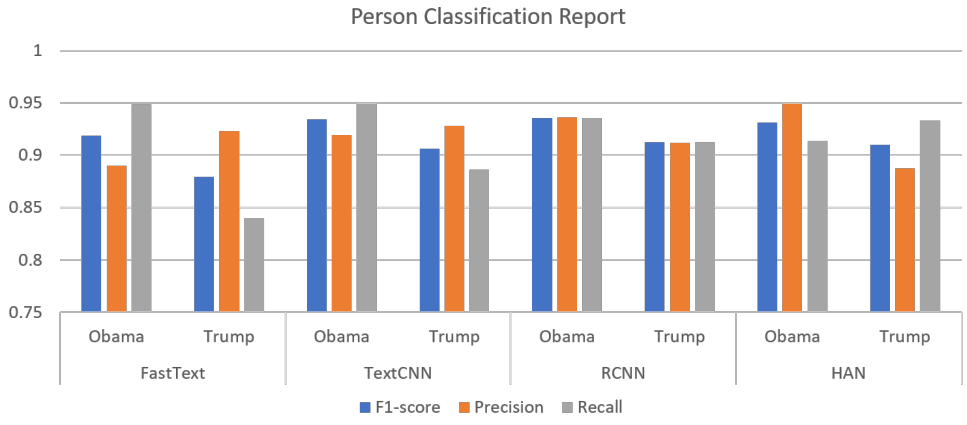
(b) TextCNN Personalization Detection



(c) RCNN Personalization Detection

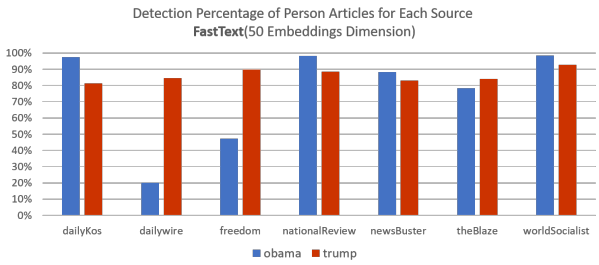


(d) HAN Personalization Detection

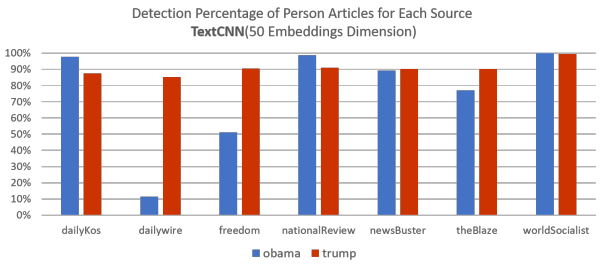


(e) FastText, RCNN, and HAN Personalization Detection Results.

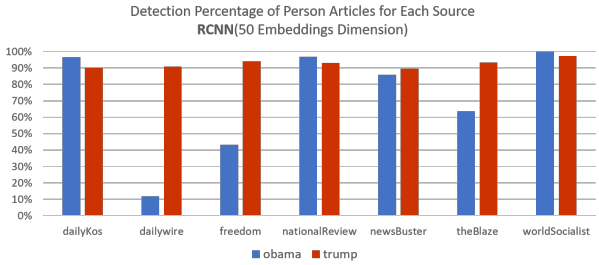
Fig. 4. Personalization Detection Comprehensive Report



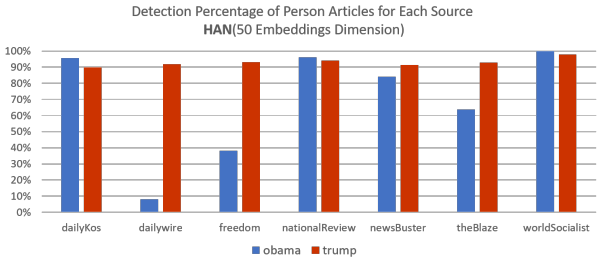
(a) Fasttext Personalization Detection Rate of News Sources



(b) TextCNN Personalization Detection Rate of News Sources

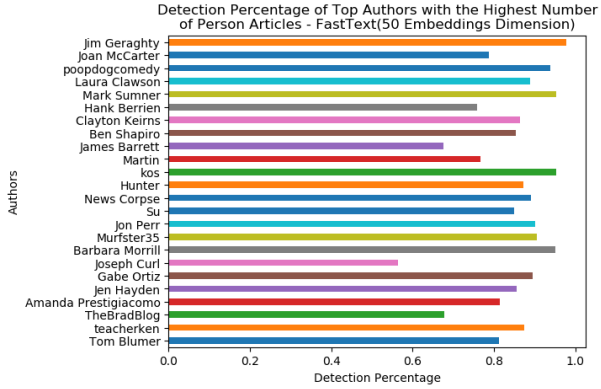


(c) RCNN Personalization Detection Rate of News Sources

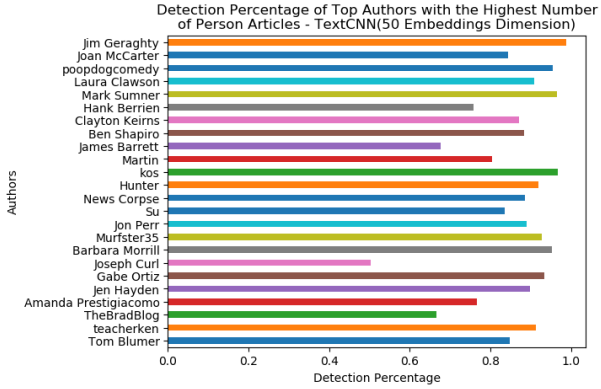


(d) HAN Personalization Detection Rate of News Sources

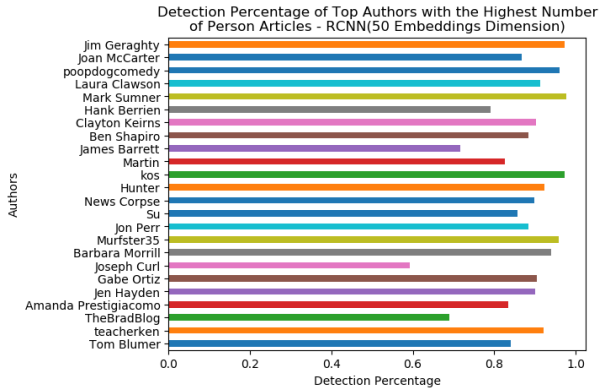
Fig. 5. Models' Personalization Detection Rate of Presidential Test Articles for All News Sources



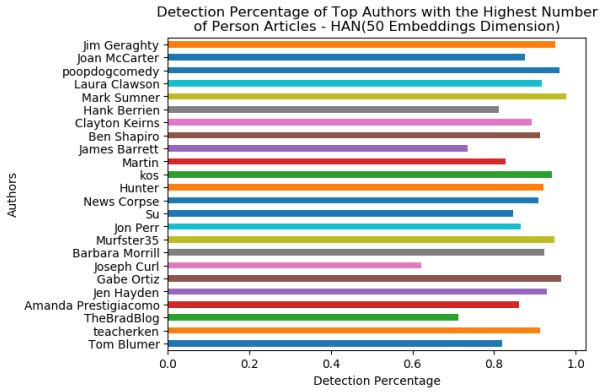
(a) FastText Personalization Detection Rate of Top Authors



(b) TextCNN Personalization Detection Rate of Top Authors



(c) RCNN Personalization Detection Rate of Top Authors



(d) HAN Personalization Detection Rate of Top Authors

Fig. 6. Models' Personalization Detection Percentage for Authors with Highest Number of Published Articles

results.

3) Politicians' Personalization in Media Sources: Observing news personalization detection percentage based on articles' sources will help identify news sources that maintain politician personalization characteristics across articles, see Fig. 5. It takes us one step further towards news framing and agenda settings detection. News politician personalization is notable in WorldSocialist as one of the far-left news media that might have a particular point of view about Trump and Obama. TextCNN, Fig. 5b, was able to accurately predict 100% of WorldSocialist's Trump and Obama articles as it has done with detecting articles' ideology. WorldSocialist received high personalization detection accuracy for both classes from the other three models with 100% for Obama and always above 92% for Trump. News article personalization does not have to be negative coverage of the politician of interest. Although National Review received non-favorable ideology detection results when labeled as Conservative, it showed that deep neural networks could identify their news articles personalization with more 90% detection accuracy. DailyKos articles have shown to be more prone to personalizing Obama articles than Trump, though that might be the result of data size skewed towards Obama articles. RCNN, HAN, and FastText (see Fig. 5c, 5d, and 5a) predicted more than 85% of Trump class articles in the remaining four conservative news websites, namely, DailyWire, IloveMyFreedom, NewsBuster, and TheBlaze, while FastText detected between 82% and 90% of Trump articles for the conservative sources. However, Obama's article prediction accuracy originated from the four conservative websites was not as impressive as Trump's articles. HAN detection percentage of politician personalized articles is as low as 9% of articles personalized on Obama and published by the DailyWire, and 20% is detected by FastText. It is often believed that data imbalance played a role in this outcome due to having more Trump articles in the three conservative sources than Obama, see Table III. Data imbalance has the same impact on detecting Obama articles originated from IloveMyFreedom, yet to a lesser degree. The gap between Trump and Obama detection accuracy is lower for NewsBusters and the Blaze news media sources. Surprisingly, there are some cases where data imbalance has no significant impact on the personalization detection models outcome, such as the DailyKos and WorldSocialist. This led us to the question of whether the poor prediction accuracy was due to data imbalance.

4) Politicians' Personalization in Authors: Authors personalize articles by shaping a persona that fits his or her point of view about the targeted politician. Unlike article ideologies, where authors compose articles that follow an ideology, news personalization is harder to define. Politically motivated authors will not easily alter their perception; hence, articles will consistently follow a pattern that might be identified. Some of the authors who had low ideology detection accuracy, got high detection accuracy for personalized articles, such as Jim Geraghty, see Fig. 6. This could mean the author has a fixated opinion about the politician in interest, but his point of view does not align with a specific ideology. The opposite is also possible where the author's articles' ideology is identified more accurately than its personalization as for TheBradBlog. Moreover, the third possible scenario where the prediction accuracy of articles' ideology and personalization of an author are both high as in Pooddogcomedy, Markos Moulitsas, and KOS. The author with the least personalized detection accuracy is Joseph Curl, with approximately 60% detection rate.

B. Detecting Articles' Political Ideology with Personalized-Based Models

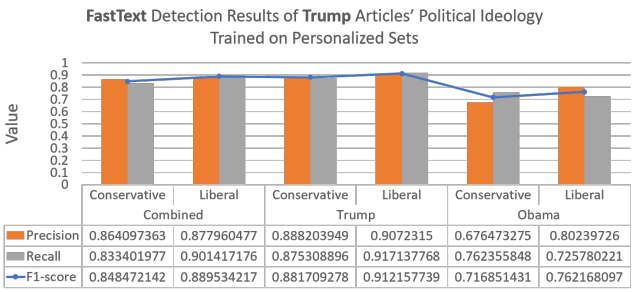
Unlike conventional linear text classifiers such SVM, deep networks require a large sum of data to deliver on various tasks, including text detection. However, learning the model on personalized articles improves or maintains ideology detection models' performance, even though the personalized training sets are a subset of the Presidential training set.

The previous statement validity is verified by training the deep neural networks on the Presidential, Trump, and Obama articles separately. Then, test the ideology detection models on Presidential, Trump, and Obama testing sets. The results obtained by those experiments are compared to each other to identify which approach is more suitable for this problem.

Fig. 7 and Fig. 8 illustrate the results of deep networks trained on the Presidential, Trump, and Obama training sets, and tested on the personalized sets only. Logically, one should expect that the accuracy of the ideology detection models will decrease when trained on a subset of the entire training set since the model will lose some information by removing a large chunk of training data. However, training on Trump set alone, Fig. 7, to detect the ideology of articles written about President Donald Trump resulted in better performance, despite removing 58% of the training set. All four deep network models scored higher on Precision, Recall, and F1-Score for Conservative and Liberal classes. Except for RCNN, Fig. 7c, that reported lower recall for the Liberal class when trained on Trump data, other metrics have increased. TextCNN, Fig. 7b, still the best ideology detection model with training on personalized data f1-score improved from 0.873 to 0.899 for the Conservative class and 0.904 to 0.924 for the Liberal class.

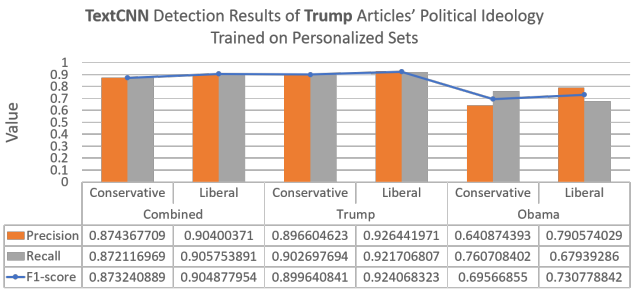
The experiment is extended to test out the ideology detection model performance when trained on Obama set to predict the ideology of articles in the Trump testing set. Although the Obama dataset is more significant in size than Trump, ideology models performed poorly compared to detection models trained on Trump or the Presidential training sets. Despite the fact these articles were collected from the same sources, the performance of models widely differs, which proved that news personalization has an impact on news article ideology detection.

Furthermore, all four deep networks retrained on Obama training set alone to predict the ideology of Obama's articles. Fig. 8a shows that FastText F1-Score slightly improved from 0.789 to 0.804 for Conservative and 0.960 to 0963 for Liberal. Other network models did not show any improvement, yet no severe decrease in their performance either. Relatively, the size of the data removed from the Presidential training set is still significant compared to information loss measure by detection model performance. Removing 42% from the entire training set did not have a severe impact on Obama's article ideology detection models. The model's performance drastically degraded when trained on Trump training set alone. The articles were collected from multiple sources with diverse topics and share



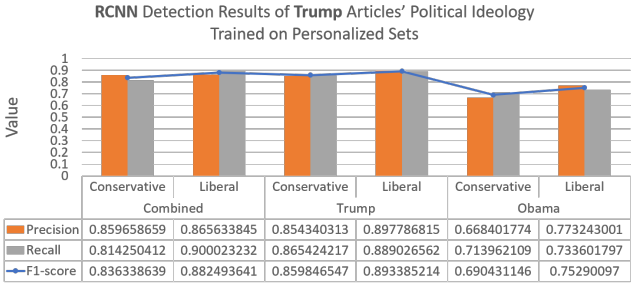
	Combined		Trump		Obama	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.864097363	0.877960477	0.888203949	0.9072315	0.676473275	0.80239726
Recall	0.833401977	0.901417176	0.875308896	0.917137768	0.762355848	0.725780221
F1-score	0.848472142	0.889534217	0.881709278	0.912157739	0.716851431	0.762168097

(a) FastText Ideology Detection Results of Trump Test Set



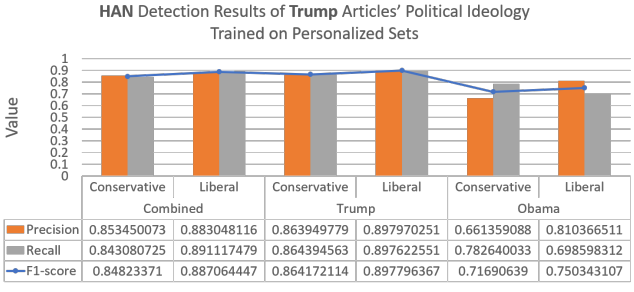
	Combined		Trump		Obama	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.874367709	0.90400371	0.896604623	0.926441971	0.640874393	0.790574029
Recall	0.872116969	0.905753891	0.902697694	0.921706807	0.760708402	0.67939286
F1-score	0.873240889	0.904877954	0.899640841	0.924068323	0.69566855	0.730778842

(b) TextCNN Ideology Detection Results of Trump Test Set



	Combined		Trump		Obama	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.859658659	0.865633845	0.854340313	0.897786815	0.668401774	0.773243001
Recall	0.814250412	0.900023232	0.865424217	0.889026562	0.713962109	0.733601797
F1-score	0.836338639	0.882493641	0.859846547	0.893385214	0.690431146	0.75290097

(c) RCNN Ideology Detection Results of Trump Test Set



	Combined		Trump		Obama	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.853450073	0.883048116	0.863949779	0.897970251	0.661359088	0.810366511
Recall	0.843080725	0.891117479	0.864394563	0.897622551	0.782640033	0.698598312
F1-score	0.84823371	0.887064447	0.864172114	0.897796367	0.71690639	0.750343107

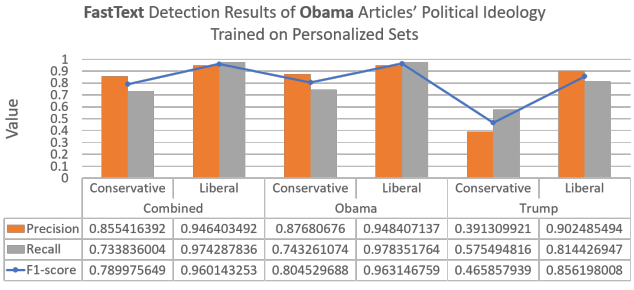
(d) HAN Ideology Detection Results of Trump Test Set

Fig. 7. Detection Results of Network Models Learned on Presidential and Personalized Training Sets to Predict the Ideology of Trump Test Articles.

a person of interest. Therefore, we believe that the results obtained from training models on the personalized sets are evidence that political personalization exists in news media articles. Reconstructing data based on personalization provides coherency and logical relation among the data, making it easier for deep networks to identify different ideological traits.

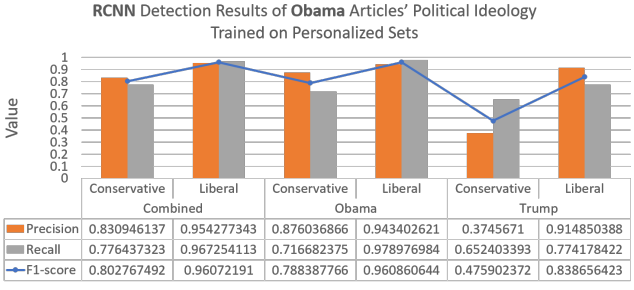
VI. Conclusion

We successfully implemented neural networks models that accurately detected politicians' personalization and political



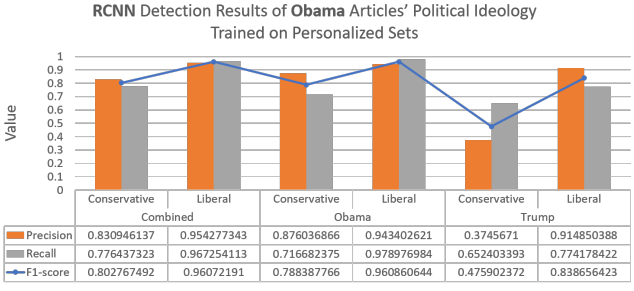
	Combined		Obama		Trump	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.855416392	0.946403492	0.87680676	0.948407137	0.391309921	0.902485494
Recall	0.733836004	0.974287836	0.743261074	0.978351764	0.575494816	0.814426947
F1-score	0.789975649	0.960143253	0.804529688	0.963146759	0.465857939	0.856198008

(a) FastText Ideology Detection Results of Obama Test Set



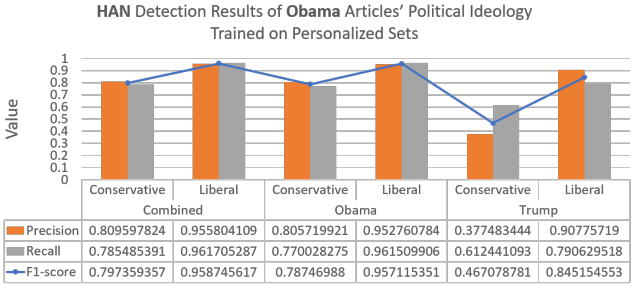
	Combined		Obama		Trump	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.830946137	0.954277343	0.876036866	0.943402621	0.3745671	0.914850388
Recall	0.776437323	0.967254113	0.716682375	0.978976984	0.652403393	0.774178422
F1-score	0.802767492	0.96072191	0.788387766	0.960860644	0.475902372	0.838656423

(b) TextCNN Ideology Detection Results of Obama Test Set



	Combined		Obama		Trump	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.830946137	0.954277343	0.876036866	0.943402621	0.3745671	0.914850388
Recall	0.776437323	0.967254113	0.716682375	0.978976984	0.652403393	0.774178422
F1-score	0.802767492	0.96072191	0.788387766	0.960860644	0.475902372	0.838656423

(c) RCNN Ideology Detection Results of Obama Test Set



	Combined		Obama		Trump	
	Conservative	Liberal	Conservative	Liberal	Conservative	Liberal
Precision	0.809597824	0.955804109	0.805719921	0.952760784	0.377483444	0.90775719
Recall	0.785485391	0.961705287	0.770028275	0.961509906	0.612441093	0.790629518
F1-score	0.797359357	0.958745617	0.78746988	0.957115351	0.467078781	0.845154553

(d) HAN Ideology Detection Results of Obama Test Set

Fig. 8. Detection Results of Network Models Learned on Presidential and Personalized Training Sets to Predict the Ideology of Obama Test Articles.

ideology in news articles, authors, and media sources. This work proved that some authors are consistent in their politicians' coverage style and more politically affiliated. Although with different degrees of bias, media sources exhibited patterns in selecting published articles. However, detecting politicians' personalization in news media is a new research topic that needs further examination. We are not aware of any research papers that studied the relation between more definitive influencing mechanisms and politicians' personalization, which will lead to new research directions that combine political science and artificial intelligence. One way to improve the work in this paper is by expanding the dataset to include more politicians or political ideologies and revaluating the detection models' performance on multiclass problems. Also, end-to-end deep neural networks can solve hierarchal problems to identify politicians' personalization and political ideology with a single network. Furthermore, other deep neural networks, such as BERT, and pre-trained networks, might achieve better results on the Presidential dataset.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4340018DSR01).

REFERENCES

[1] J. Dunaway, K. Searles, M. Sui, and N. Paul, "News attention in a mobile era," *Journal of Computer-Mediated Communication*, vol. 23, no. 2, pp. 107–124, 2018.

[2] S. Schäfer, "Illusion of knowledge through facebook news? effects of snack news in a news feed on perceived knowledge, attitude strength, and willingness for discussions," *Computers in Human Behavior*, vol. 103, pp. 1–12, 2020.

[3] H. Wen, L. Fang, and L. Guan, "A hybrid approach for personalized recommendation of news on the web," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5806–5814, 2012.

[4] H. J. Lee and S. J. Park, "Moners: A news recommender for the mobile web," *Expert Systems with Applications*, vol. 32, no. 1, pp. 143–150, 2007.

[5] R. C. Bagher, H. Hassanpour, and H. Mashayekhi, "User trends modeling for a content-based recommender system," *Expert Systems with Applications*, vol. 87, pp. 209–219, 2017.

[6] A. Montes-García, J. M. Álvarez-Rodríguez, J. E. Labra-Gayo, and M. Martínez-Merino, "Towards a journalist-based news recommendation system: The wesomender approach," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6735–6741, 2013.

[7] K. Thorson, K. Cotter, M. Medeiros, and C. Pak, "Algorithmic inference, political interest, and exposure to news and politics on facebook," *Information, Communication & Society*, pp. 1–18, 2019.

[8] G. de Souza Pereira Moreira, F. Ferreira, and A. M. da Cunha, "News session-based recommendations using deep neural networks," in *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, 2018, pp. 15–23.

[9] V. Kumar, D. Khattar, S. Gupta, M. Gupta, and V. Varma, "Deep neural architecture for news recommendation." in *CLEF (Working Notes)*, 2017.

[10] C. Chen, X. Meng, Z. Xu, and T. Lukasiewicz, "Location-aware personalized news recommendation with deep semantic analysis," *IEEE Access*, vol. 5, pp. 1624–1638, 2017.

[11] I. Dylko, I. Dolgov, W. Hoffman, N. Eckhart, M. Molina, and O. Aaziz, "The dark side of technology: An experimental investigation of the influence of customizability technology on online political selective exposure," *Computers in Human Behavior*, vol. 73, pp. 181–190, 2017.

[12] N. M. Anspach, "The new personal influence: How our facebook friends influence the news we read," *Political Communication*, vol. 34, no. 4, pp. 590–606, 2017.

[13] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Social influence and unfollowing accelerate the emergence of echo chambers," *Journal of Computational Social Science*, pp. 1–22, 2020.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[15] C. Qu, F. Ji, M. Qiu, L. Yang, Z. Min, H. Chen, J. Huang, and W. B. Croft, "Learning to selectively transfer: Reinforced transfer learning for deep text matching," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 699–707.

[16] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," *arXiv preprint arXiv:1704.05742*, 2017.

[17] J. T. Hamilton and R. Zeckhauser, "Media coverage of ceos: who? what? where? when? why?" *Unpublished working paper, Stanford Institute of International Studies*, 2004.

[18] F. Bendisch, G. Larsen, and M. Trueman, "Fame and fortune: a conceptual model of ceo brands," *European Journal of Marketing*, 2013.

[19] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, pp. 1–27, 2019.

[20] O. Tsur, D. Calacci, and D. Lazer, "A frame of mind: Using statistical models for detection of framing and agenda setting campaigns," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1629–1638.

[21] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding," *Expert Systems with Applications*, vol. 139, p. 112851, 2020.

[22] F. Enríquez, J. A. Troyano, and T. López-Solaz, "An approach to the use of word embeddings in an opinion classification task," *Expert Systems with Applications*, vol. 66, pp. 1–6, 2016.

[23] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: an interdisciplinary literature review," *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 391–415, 2019.

[24] K. M. Alzhrani, "Political ideology detection of news articles using deep neural networks," *Intelligent Automation & Soft Computing*, vol. 33, no. 1, pp. 483–500, 2022.

[25] D. Ball, P. S. Coelho, and M. J. Vilares, "Service personalization and loyalty," *Journal of services marketing*, 2006.

[26] W. Meng, R. Ding, S. P. Chung, S. Han, and W. Lee, "The price of free: Privacy leakage in personalized mobile in-apps ads." in *NDSS*, 2016.

[27] F. Essalmi, L. J. B. Ayed, M. Jemni, S. Graf *et al.*, "A fully personalization strategy of e-learning scenarios," *Computers in Human Behavior*, vol. 26, no. 4, pp. 581–591, 2010.

[28] O. Ben-Shahar and A. Porat, "Personalizing negligence law," *NYUL Rev.*, vol. 91, p. 627, 2016.

[29] J. Vesanen, "What is personalization? a conceptual framework," *European Journal of Marketing*, 2007.

[30] H. Fan and M. S. Poole, "What is personalization? perspectives on the design and implementation of personalization in information systems," *Journal of Organizational Computing and Electronic Commerce*, vol. 16, no. 3-4, pp. 179–202, 2006.

[31] B. Maddens and S. Fiers, "The direct pm election and the institutional presidentialisation of parliamentary systems," *Electoral Studies*, vol. 23, no. 4, pp. 769–793, 2004.

[32] D. R. Kinder, M. D. Peters, R. P. Abelson, and S. T. Fiske, "Presidential prototypes," *Political behavior*, vol. 2, no. 4, pp. 315–337, 1980.

[33] F. F. da Silva, D. Garzia, and A. De Angelis, "From party to leader mobilization? the personalization of voter turnout," *Party Politics*, 2019.

[34] T. Bøggild and H. H. Pedersen, "Campaigning on behalf of the party? party constraints on candidate campaign personalisation," *European Journal of Political Research*, vol. 57, no. 4, pp. 883–899, 2018.

[35] S. Kruikemeier, G. Van Noort, R. Vliegenthart, and C. H. De Vreese, "Getting closer: The effects of personalized and interactive online political communication," *European journal of communication*, vol. 28, no. 1, pp. 53–66, 2013.

[36] I. McAllister *et al.*, "The personalization of politics," in *The Oxford handbook of political behavior*. Oxford University Press, 2007.

[37] S. Mainwaring and M. Torcal, "Party system institutionalization and party system theory after the third wave of democratization," *Handbook of party politics*, vol. 11, no. 6, pp. 204–227, 2006.

[38] A. Mughan, *Media and the presidentialization of parliamentary elections*. Springer, 2000.

[39] T. Poguntke and P. Webb, *The presidentialization of politics: A comparative study of modern democracies*. Oxford University Press on Demand, 2007.

[40] ——, "The presidentialization of politics in democratic societies: A framework for analysis," *The presidentialization of politics: a comparative study of modern democracies*, vol. 1, 2005.

[41] P. Webb and T. Poguntke, "The presidentialisation of politics thesis defended," *Parliamentary Affairs*, vol. 66, no. 3, pp. 646–654, 2013.

[42] K. Dowding, "The prime ministerialisation of the british prime minister," *Parliamentary Affairs*, vol. 66, no. 3, pp. 617–635, 2013.

[43] D. Garzia, *Personalization of politics and electoral change*. Springer, 2019.

[44] H. Kriesi, "The populist challenge," *West European Politics*, vol. 37, no. 2, pp. 361–378, 2014.

[45] M. Canovan, "Trust the people! populism and the two faces of democracy," *Political studies*, vol. 47, no. 1, pp. 2–16, 1999.

[46] A. I. Langer and I. Sagarzazu, "Bring back the party: personalisation, the media and coalition politics," *West European Politics*, vol. 41, no. 2, pp. 472–495, 2018.

[47] A. Bittner, *Platform or personality?: the role of party leaders in elections*. OUP Oxford, 2011.

[48] L. Karvonen, *The personalisation of politics: A study of parliamentary democracies*. Ecpr Press, 2010.

[49] M. Kaase, "Is there personalization in politics? candidates and voting behavior in Germany," *International Political Science Review*, vol. 15, no. 3, pp. 211–230, 1994.

[50] P. Van Aelst, T. Sheafer, and J. Stanyer, "The personalization of mediated political communication: A review of concepts, operationalizations and key findings," *Journalism*, vol. 13, no. 2, pp. 203–220, 2012.

[51] M. Balmas, G. Rahat, T. Sheafer, and S. R. Shenhav, "Two routes to personalized politics: Centralized and decentralized personalization," *Party Politics*, vol. 20, no. 1, pp. 37–51, 2014.

[52] M. Balmas and T. Sheafer, "Leaders first, countries after: Mediated political personalization in the international arena," *Journal of communication*, vol. 63, no. 3, pp. 454–475, 2013.

[53] J. Takens, J. Kleinnijenhuis, A. Van Hoof, and W. Van Atteveldt, "Party leaders in the media and voting behavior: Priming rather than learning or projection," *Political Communication*, vol. 32, no. 2, pp. 249–267, 2015.

[54] B. Wauters, P. Thijssen, P. Van Aelst, and J.-B. Pilet, "Centralized personalization at the expense of decentralized personalization. the decline of preferential voting in belgium (2003–2014)," *Party Politics*, vol. 24, no. 5, pp. 511–523, 2018.

[55] G. Rahat and T. Sheafer, "The personalization (s) of politics: Israel, 1949–2003," *Political communication*, vol. 24, no. 1, pp. 65–80, 2007.

[56] S. Adam and M. Maier, "Personalization of politics a critical review and agenda for research," *Annals of the International Communication Association*, vol. 34, no. 1, pp. 213–257, 2010.

[57] K. Alzhrani, "Ideology detection of personalized political news coverage: A new dataset," in *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, 2020, pp. 10–15.

[58] Y. Li, G. Sun, and Y. Zhu, "Data imbalance problem in text classification," in *2010 Third International Symposium on Information Processing*. IEEE, 2010, pp. 301–305.

[59] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[60] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[61] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[62] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

Multi-Spectral Imaging for Fruits and Vegetables

Shilpa Gaikwad*
Electronics and Telecommunication Engineering
Symbiosis Institute of Technology
Symbiosis International University
Pune, India, 412115

Sonali Tidke
Computer Science and Engineering
Symbiosis Institute of Technology
Symbiosis International University
Pune, India, 412115

***Abstract*—In the field of agriculture, fruit grading and vegetable classification is an important and challenging task. The current fruit and vegetable classifications are done manually, which results in inconsistent performance. There is an influence of external surroundings on this manual classification. Sometimes getting an expert fruit or vegetable grader, is challenging and the performance of that person may become stagnant over some time. With the recent development in computer technology and multi-spectral camera system, it is possible to achieve an efficient fruit grading or vegetable classification system. In this manuscript, we summarize different automated fruit grading as well as vegetable classification systems, which are based on multi-spectral imaging. We have focused our analysis on four major areas such as varietal identification, fruit quality assessment, classification, and disease detection. From our analysis, we have found that the Partial Least Square Discriminant Analysis (PLS-DA) was most effective for identifying varieties of tomato seeds. For analyzing the quality of pomegranate fruits, the multiple linear regression model was the most optimal method. Multi-Layer Perceptron (MLP) classifier was the recommended method for classifying medicinal plant leaves. A Linear Discriminant Analysis (LDA) based classifier could accurately detect diseases in fruits and vegetables.**

***Keywords*—Multi-spectral imaging; fruit grading; vegetable classification; fruit quality; disease detection; fruit maturity**

I. Introduction

A multi-spectral image is the set of images for different wavelengths, which is captured throughout the electromagnetic spectrum [2]. Wavelength separation for image data is typically achieved with the help of color filters [3]. Multi-spectral imaging not only includes visible wavelength but sometimes the wavelength outside the visible spectrum (i.e. infrared and ultra-violet). With the recent advancement in manufacturing techniques, Light Emitting Diodes (LED's) with multiple colors such as red, green, blue (RGB) came into the picture and they became very popular [4]. The information that we cannot visualize with the human eye, can be visualized with the help of multi-spectral imaging. In multi-spectral imaging, there are at least two to five spectral bands involved. A multi-spectral system consists of the following spectral bands: visible spectra are in the range of 0.4 to 0.7 um. The range of near-infrared spectra (NIR) is 0.7 to 1 um, the short wave infrared (SWIR) is visible in the range of 1 to 1.7 um, the mid-wave infrared (MWIR) is visible in the range of 3.5 to 5 um and long-wave infrared (LWIR) is visible in the range of 8 to 12 um. These are combined into a single system [5]. Multi-spectral imaging can be done with a minimum of 3 and a maximum of 15 spectral bands. The applications of multi-spectral imaging include: detecting or tracking military targets, detecting landmines, detection of a ballistic missile, weather forecasting, space-based imaging, and investigation of paintings and documents [6]. Fig. 1 shows the schematic of multi-spectral imaging. The light consisting of a large range of wavelengths (including visible and non-visible) is passed through a multi-spectral filtered disk to generate a specific wavelength of light. A fruit sample is kept into a specific wavelength and a camera captures the image of the sample. Now the multi-spectral filter disk is rotated to obtain other slices of the multi-spectral image. Finally, the computer is used to process the multi-spectral image with the help of algorithms to obtain the grading of the fruit/vegetable sample. The camera doesn't have any IR filter attached to it so that all the wavelengths of light can be observed. Fig. 2 shows the overview of multi-spectral imaging for various applications in fruits and vegetables.

A. Benefits of Employing Multi-Spectral Imaging for Analysing Fruits and Vegetables

By using multi-spectral imaging for analyzing fruits and vegetables, a lot of important information about the fruits can be extracted. The primary use of multi-spectral imaging in the context of fruits and vegetable analysis is to determine the quality of the product in a non-invasive manner. Higher quality fruits can be marked for export to foreign countries. A particular country can establish its reputation for exporting high-quality fruits and vegetables. In the domestic markets, high-quality fruits and vegetables can be sold at a premium price hence motivating farmers to cultivate a better quality of crops. Another application of multi-spectral imaging is to determine the sugar content of fruits before they are ripe. Assessing the sugar content can help in achieving higher prices for fruits that have high sugar content. Each species of fruit and vegetable would generate unique multi-spectral images. Hence, such imaging techniques would help in identifying the breed of the crop so that it can be properly graded. By analyzing such images of a small number of crops in a field, the eventual overall yield of fruits and vegetables from that field can also be predicted. Multi-spectral imaging is also used to identify disease-stricken fruits and vegetables in quality control systems. In this manuscript, we evaluate the various applications of multi-spectral imaging and remark on the most optimal methods of analyzing such images for each area of application. Main areas of application of multi-spectral imaging are:

1. Varietal identification
2. Fruit quality analysis
3. Classification of fruits

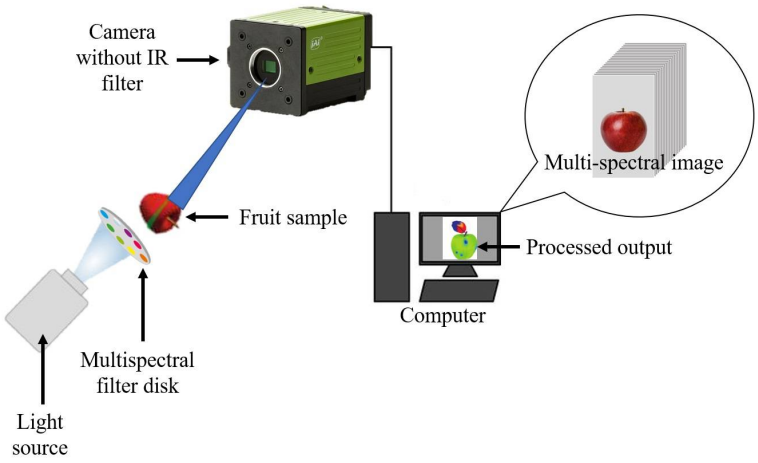


Fig. 1. Schematic Diagram of Multi-Spectral Imaging. The Light Source Wavelengths (Including Visible and Non-Visible) are Passed to a Multi-Spectral Filtered Disk to Generate a Specific Wavelength of Light. A Fruit Sample is Kept at a Specific Wavelength and a Camera Captures the Image of the Sample. Now the Multi-Spectral Filter Disk is Rotated to Obtain Other Slices of the Multi-Spectral Image. Finally, the Computer is used to Process the Multi-Spectral Image with the Help of Algorithms to Obtain the Grading of the Fruit/Vegetable Sample. The Camera used does not have any IR Filter so that All Wavelengths of Light can be Observed.

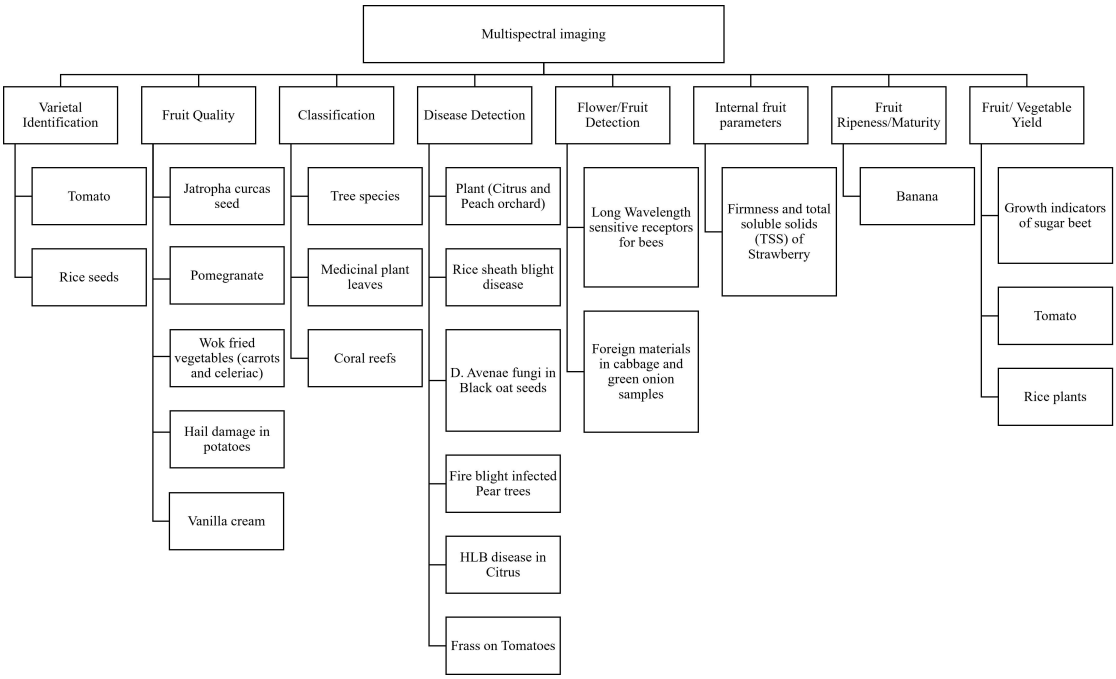


Fig. 2. Overview of Applications of Multi-Spectral Imaging related to the Analysis of Fruits and Vegetables. Varietal Identification is Generally used with Tomato and Rice Seeds. The Quality of Pomegranate Fruits and Jatropha Curcas Seeds can be determined by using Multi-Spectral Imaging. Also, Assessment of Hail Damage Analysis and Differentiation of Vanilla Cream Samples into Fresh and Spoiled Classes can be Performed. It can be used for Classifying Medicinal Plant Leaves, Tree Species and Coral Reefs. Fruits and Flowers can be Identified based on their Multi-Spectral Images. Internal Fruit Parameters are used to Calculate Total Soluble Solids and Firmness in Fruits are measured. The Fruit Ripeness or Maturity Detection can be Performed for Bananas. Another Application of Multi-Spectral Imaging is Fruit and Vegetable Yield Prediction.

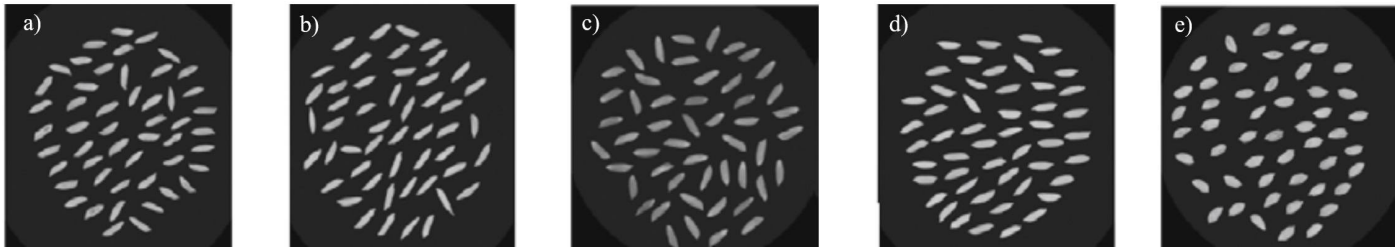


Fig. 3. (a-e) Rice Seeds Classification. Multi-Spectral Imaging Enables the Rice to be Classified as FD2 (a), QXY512 (b), HXD3 (c), QXY822 (d), and WKJ11 (e). These Five Types are Varietal Identification Numbers of Rice [1].

4. Disease detection
5. Flower or fruit detection
6. Internal fruit parameter extraction
7. Fruit ripeness / maturity estimation
8. Fruit Yield

II. Literature Review

Fruits and vegetables are very important sources of our daily nutrition. From a consumer's point of view, the quality of fruits and vegetables is a vital feature. Fruit and vegetable quality include both internal and external properties. The internal quality is mainly determined by smell, taste, texture, firmness of flesh, diseases, organic residues, and nutritional quality which includes sugar content, starch, organic acids, total soluble solids content, etc. The external quality is mainly determined by shape, size, color, appearance, and bruises [7].

A. Varietal Identification

Quality and quantity of yield production are estimated by using a varietal identification-based techniques. Liu *et al.* [1] used a combination of multi-spectral imaging and chemometric data to find the varieties of rice seeds in a fast manner without damaging the rice seeds. In this study, multi-spectral imaging methods were applied for the separation of rice seeds based on different varieties. Morphological and spectral features were extracted from multi-spectral imaging. Different chemometric methods along with the Partial Least Squares Discriminant Analysis (PLS-DA), Least Squares-Support Vector Machine (LS-SVM) models, and Principal Component Analysis-Back Propagation Neural Network (PCA-BPNN) were used. In the next step, their discrimination performance was compared to classify the rice seeds into five different varieties. The spectral data consisted of various rice seed features. Lastly, the spectral and morphological information were combined respectively and the discrimination performance was compared. After comparing, some differences were found in the varieties of rice seeds and they could be classified according to their variety as shown in (Fig. 3(a-e)). The classification accuracy was up to 94% for the LS-SVM model, 62% for the PLS-DA, and 84% for the PCA-BPNN models.

Recently, Shrestha *et al.* [8] has proposed a method of varietal separation and recognition of tomato cultivars using Principal Component Analysis (PCA) along with Normalized Canonical Discriminant Analysis (nCDA). They aimed to calculate and compare the parents' and next-generation results obtained by using the method. They considered two sample sets out of which the first set comprised of two cultivars and their two crosses for studying parent and offspring affinity. 11 cultivars were used in the second sample set for varietal identification. With the help of a VideometerLab instrument, multi-spectral images were captured. A blob database was constructed with images of all seeds. Blob feature RegionMSI (mean) could achieve the best separation when it was compared with the intensity of the pixel (mean). The calibration set values are compared with the results obtained from the unknown set. The model could accurately identify the different varieties with an accuracy of 82 %. For step-wise PLS-DA models, the classification error was 7 % for the cross-validation dataset as well as the test dataset.

1) Varietal Identification Setup: Images from each seed sample (tomato or rice) are captured using a VideometerLab instrument as shown in Fig. 4. This setup captured multi-spectral images in 19 different wavelengths such as 375,405, 435, 450, 470, 505, 525, 570, 590, 630, 645, 660, 700, 780, 850, 870, 890, 940 and 970 nm [8]. It consists of a sphere that has a white coating to take care of the light getting uniformly scattered around the object. The sphere consists of 19 LEDs along its rim along with a camera that is mounted at the top. Initially, a standard target was used to calibrate the system radiometrically and geometrically. A light setup was employed depending on the type of the sample to make sure that direct comparable images were captured [1]. The background information in the captured image is not relevant hence, a CDA method was used to remove it and only retain the Region of Interest (ROI). The VideometerLab software was used to extract data and transform the pixel data once the images were captured. Various algorithms such as n-CDA, PCA, PLS-DA, BPCNN, and LS-SVM were used. The seed images were segmented based on a certain threshold value. Morphological features such as the area (mm^2), width/length, and roundness values of seeds were extracted for analysis of the image [1].

B. Fruit Quality

The quality of fruit is a vital factor when considered from point of view of the consumers. Human experts can examine the quality of the fruit [7]. But manual sorting by visual inspection is work-intensive, slow, inconsistent, and incorrect. With the invention of multi-spectral imaging techniques, automation of the grading process can be performed which will reduce labor costs, and increase the efficiency and accuracy of the sorting process. Fruit quality includes both internal and external properties. The internal quality of fruits includes taste, texture, smell, flavor, flesh firmness, diseases,

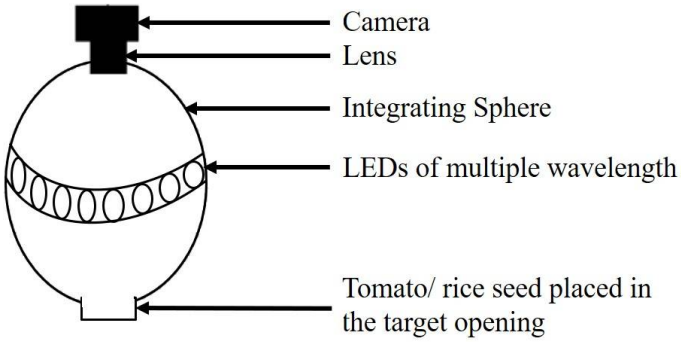


Fig. 4. Structural Setup for Capturing Multispectral Images using the VideometerLab Instrument [8].

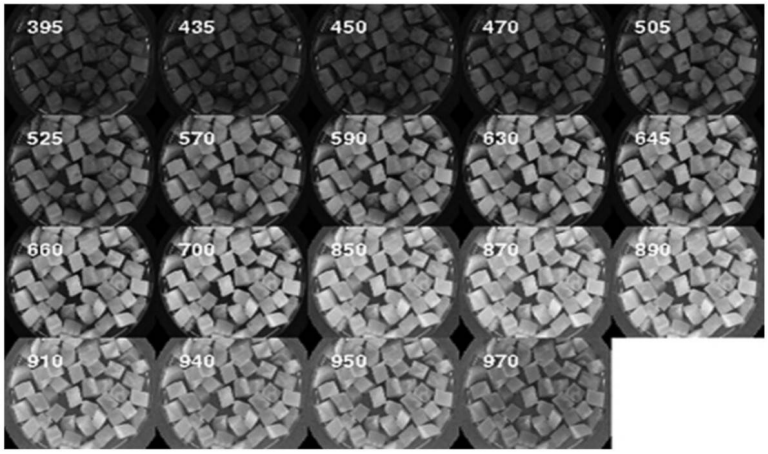


Fig. 5. Plot showing All Spectral Channels of a Celeriac Sample [9].

and chemical/organic residues. It also consists of nutritional quality which includes sugar content, starch, organic acids, total soluble solids content, etc. The quality of fruit depends on the shape, size, color, and skin defects (bruises). The maturity and skin defects are the most crucial features that are used to estimate fruit quality. To evaluate the quality of fruits, internal fruit parameters, ripeness, and yield should be analyzed. An automatic fruit and vegetable grading system will help both farmers and consumers by providing high-quality fruits in the market. Khodabakhshian *et al.* [10] has developed a multi-spectral imaging system to assess pomegranate fruit quality. A visible/NIR spectroscopy in the 400–1100 nm range was used for finding the Total Soluble Solids (TSS), pH, and titratable acidity (TA). The performance of a multi-spectral imaging system was estimated using a multiple linear regression model. The resultant TSS has the value of correlation coefficient (r) as 0.97, the Root Mean Square Error of calibration (RMSEC) was 0.21°Brix and the Ratio Performance Deviation (RPD) is 6.7 °Brix. Hence, it is shown that the models have a great ability to predict pH and TA.

1) Block Diagram of Pomegranate Fruit Quality: **Spectrometer:** As shown in Fig.6 (a) the spectrometer is provided with an optical fibre cable externally. To capture the Vis/NIR diffused reflectance spectra from the pomegranate fruit the measurement system is operated in reflectance mode. For each fruit, an average value of a total of 30 scans was taken. **Evaluation of TSS, TA, and pH:** Once the spectra were acquired, the TSS, TA, and pH of each fruit sample were calculated. A manual fruit squeezer was used to find the above-mentioned features. This was followed by the filtration and centrifugation of the fruit. The TSS was measured in °Brix three times with the help of a hand-held refractometer. Similarly, a digital pH meter was used to measure the pH. The average of these measured values for both parameters was recorded. An average percentage of the citric acid was measured after the value of TA was acquired with the help of a Metrohm 862 compact titro sampler. **Preprocessing:** A huge amount of spectral data is produced with the help of Vis/NIR instruments. However, there is some excessive data captured such as the background noise in the data acquired from the spectrometer apart from the required information on samples. The effect of this irrelevant information is reduced by pre-processing the data. This is done to achieve accurate, stable, and reliable calibration models. The pre-processing techniques utilized for this purpose can be categorized as columns pretreatments and rows pretreatments. ParLeS software was employed to implement the pretreatments. Initially, the average of four spectra was taken to get a single spectrum. An absorbance value was achieved by converting this spectrum using the equation $Abs = log(1/R)$ where R is the quantity

of reflectance. A linear correlation was achieved between the molecular concentration of the sample and the spectra. Lastly, various methods for pre-processing such as centering, normalization, smoothing, and differentiation were performed. Centering was used to ensure the best results in terms of change across the mean. Smoothing was used to find the best Signal to Noise (SNR) ratio. Multiplicative Scatter Correction (MSC) was used to eliminate the results of scattering related to the average spectrum. Similarly, the background spectra were eliminated and the spectral resolution was increased by using the 1st and 2nd derivative pre-processing methods. **Calibration and Validation:** A model was developed using the PLS regression method between quality parameters (TA, TSS, and pH) from samples and spectral responses of the samples. ParLeS software was used for comparing one quality parameter and the spectral data. The calibration set consisted of 70 samples. The purpose of the PLS method is to find a mathematical relationship between a set of independent variables/predictors which consists of the X matrix (70 fruits × 700 wavelengths) and the dependent variable which consists of the Y matrix (70 fruits × 1). The dependent variables (Y) are defined by the quality parameter (TSS, TA, and pH) value from the calibration set. For validation 30 fruits were selected. When the predicted residual error sum of squares was minimum, the calibration model was built using the maximum number of latent elements. The correlation coefficient (r), Root Mean Square Error of Calibration (RMSEC), Root Mean Square Error of Prediction (RMSEP), and Ratio Performance Deviation (RPD) were measured to evaluate the results of the validation and calibration model. Clemmensen *et al.* [9] proposed an imaging technique for capturing high-quality multi-spectral images in carrots and celeriac for 14 days to observe optical reflection changes. For this purpose, the vegetables were fried and frozen at 30^0C for 4 months before recording a multi-spectral image. Fig. 5 shows the plot of all spectral channels of a celeriac sample. The vegetables were kept at $+5^0$C throughout the image capturing phase of 14 days. During this period, surface changes, as well as the reflectance properties, were very subtle. But, numerically important differences for some wavelengths and combinations of wavelengths were observed. A t-test was conducted to check for important differences on a 5 % level in many percentage points of the light reflectance. Major variations were observed in the reflectance spectrum of the carrots and celeriac from days 2 to 4. While for the celeriac, significant changes in the reflectance spectra till day 14 were noted. Bhargava *et al.* [11] has proposed a computer vision-based method to analyze quality of apples. They have used image processing techniques on the input apple images, along with an Artificial Neural Network (ANN) to achieve 96 % accuracy in quality assessment.

2) Block Diagram of Wok-Fried Vegetables to Find Optical Reflection Changes:: The VideometerLab setup was used to capture the multi-spectral images refer Fig. 6(c). T-tests were performed on the multispectral images to find the importance of reflection changes between days of the experiment. Pixel intensity values of the features were measured at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for each vegetable piece kept in the refrigerator. This was done to examine the reflection changes as a function of days. To categorize the differences 358 carrot samples and 389 celeriac samples were tested in the interval of six different sample days. The p-values for the T-tests were plotted with the null hypothesis and it showed no variation in mean values for the six-day interval. Lianou *et al.* [12] proposed an online feature selection grading scheme for vanilla cream quality analysis using multi-spectral imaging. The study was conducted for the inspection of two microbiological quality classes of cream samples, with the value of total viable counts (TVC) $\leq$ 2.0 log CFU/g for fresh samples and TVC $\geq$ 6.0 log CFU/g for spoiled samples. They achieved an overall classification accuracy of 91.7 %. This model can be further extended to find the microbiological quality of classes by using 1 log step. This step was taken to validate the capability of the model to evaluate increasing microbial populations.

3) Block Diagram to Find Vanilla Cream Quality: **Samples of vanilla cream subjected to microbiological study** Fig. 6(c) shows the block diagram of the vanilla cream sample. The vanilla cream sample of 25gm weight was kept in a sterile stomacher bag having 225 mL of the sterilized 1/4th concentration solution of Ringer. It was homogenized in a Stomacher device at room temperature for 1 min. To calculate the TVC, precise serial decimal dilutions in Ringer's solution were considered. Finally, colonies were measured after an incubation period of 48 hours with the temperature maintained at 30^oC. The microbiological quality of data was indicated as log (colony forming units) per gram of cream (log CFU/g). **Image capturing and its study:** The VideometerLab setup was used to capture the multi-spectral images for each cream sample. Capturing the spectral data, a data cube of m x n size in pixels is captured for each cream sample. This resulted in the production of a huge amount of data describing samples in time-based experiments considering different storage conditions. Before the image was captured a light setup called auto light was used and the calibration was performed radiometrically and geometrically. The samples of vanilla cream were placed in an Ulbricht sphere which has a camera installed on the top and the subsequent multispectral image of the cream surface was captured. To eliminate the redundant data, segmentation was done using the in-house method. After this normalization was done for pixel values of each wavelength in the range [0,1]. The next step was to calculate the mean reflectance spectrum and the respective standard deviation values were calculated. **Data labelling:** Out of the 245 spectra of the cream samples, 65 spectra were used in model training, 48 spectra were used for validating the model and the remaining 132 spectra were used in model testing. The input matrix had 18 average and 18 Standard Deviation (SD) values of the reflective spectra extracted from 245 samples of vanilla cream. The output matrix had the results of microbiological data for finding TVCs in the corresponding samples. At first, binary classification was applied that will automatically find samples having TVC $\geq$ 6 log CFU/g and TVC $\leq$ 2 log CFU/g. The model was constructed to correctly find fresh or spoiled samples. **Dynamic Feature Selection (DFS) method:** An Unsupervised Online Feature Selection (UOS) algorithm based on the DFS method was used to choose the best model features corresponding to certain test data used. A three-step DFS approach was designed. The first step was training-dependent feature elimination. The second step was the online test-dependent feature elimination. The third step was the decision taken by considering the training set and test

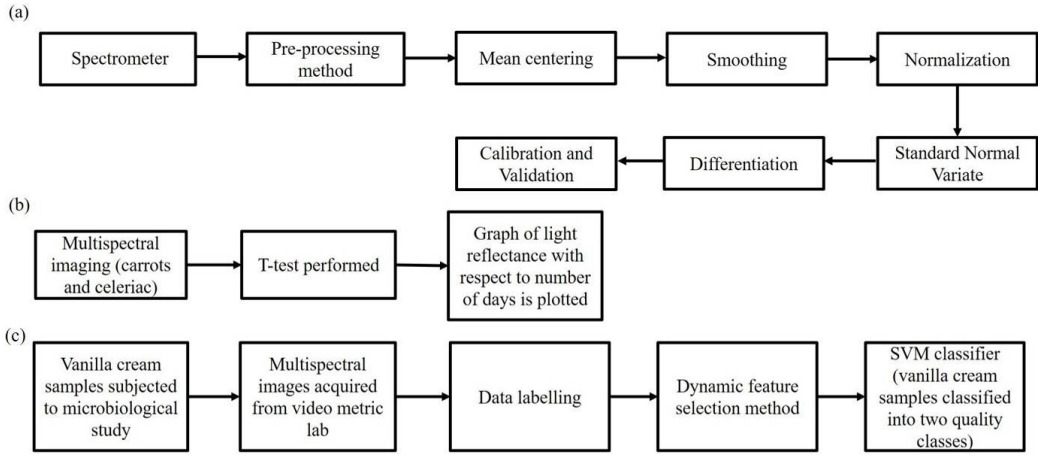


Fig. 6. Block Diagram for Fruit Quality. (a) Chemometric Block Diagram to Extract the Information of Pomegranate Fruit. (b) Block Diagram to Calculate Optical Reflection Changes in Carrots and Celeriac using Multispectral Imaging. (c) Block Diagram of SVM Classifier to Classify Vanilla Cream Samples in Two Quality Classes.

samples collected. The classification model is built when the training set features fit all three conditions for each test sample. **Support Vector Machine (SVM) classifier:** SVM with linear kernel was used to classify the vanilla cream samples. Every new test sample produces a new set of features. These features are re-inserted in the data set for training and certain features were selected to create new test data in the online feature selection method.

4) Internal Fruit Parameters: Liu *et al.* [13] proposed the use of multi-spectral Imaging to find quality parameters and ripeness phases of the strawberry fruit. A total of 210 fruits were evaluated. For developing models, the calibration set included 162 fruits which comprised of raw, ripe, and overripe classes with 54 samples each. For model validation, 48 fruits were considered which comprised of raw, ripe, and overripe samples with 16 samples each. This was done so that the calibration model could verify the prediction performance. The results show that for predicting firmness and TSS content, the BPNN model's performance was better as compared with PLS and SVM models using multi-spectral imaging. The SVM and PCA-BPNN models were tested to classify fruit based on the maturity stage with the help of multi-spectral imaging. The SVM model achieved a classification accuracy of 100 %.

5) Block Diagram to Find Internal Fruit Parameters (TSS and Firmness) in Strawberry Fruit: **Sample preparation and multi-spectral imaging:** 210 strawberry fruits were considered (70 raw and 40 ripe samples) (Fig. 13). Multi-spectral images were captured for all the fruits. To generate the overripe fruit category, the ripe fruits were kept at room temperature for two days. A texture analyzer was used to perform penetration tests on the skin of the fruit with a depth of 7 mm. The firmness of the fruit was also calculated. The TSS was measured with a handheld refractometer. The measurement of each fruit was done thrice immediately after the multi-spectral image measurement. The VideometricLab equipment was used for capturing multi-spectral images of the strawberry fruit. Segmentation of the images into proper regions was performed for image evaluation. The nCDA algorithm was used to remove the background information and segmented using a simple threshold. Otsu adaptive thresholding method was used to emphasize the required features which were followed by segmentation. The images of strawberries after removal of the background were transformed into spectra depending on the calculations of the mean. Each image was contributing its spectra for model calibration. **Data evaluation:** The model was built by considering the acquired spectra and the quality features like TSS and firmness. The SVM, PLS, and BPNN methods were used to develop the prediction model.

6) Fruit Ripeness or Maturity: With the help of multi-spectral imaging, a method for estimating the ripeness of bananas is proposed by Santoyo *et al.* [14]. The brown spots on the banana peel can be separated accurately by employing the Hotelling transform. Two optical filters that were visible in the range of 410-690 nm and NIR in the range of 820-910 nm were used for this purpose. The texture homogeneity criteria were used to specify the growth of the brown spots present in process of ripening with the fusion of spectral image. Fig. 7 shows the 7 phases of the banana ripening process.

7) Block Diagram to Find Banana Ripeness: In this work, seven specimens of bananas were taken. Each day 10 images were taken for every specimen (one for every optical filter) (as shown in Fig. 8). This process resulted in 490 images. The banana specimens were stored for 7 days. 10 different optical filters were used in this experiment. The flat-field correction technique was applied to remove the unwanted signals and retain the useful data for each multi-spectral image. To estimate the maturity of the banana, quantification of the brown spots growing on the peel was carried out. Three quantifying methods namely Fourier fractal analysis, Hotelling transform, and texture analysis were used. These techniques were employed to compute the homogeneity criteria related to the co-occurrence matrix. These methods also helped to find the level of similarity in a set of pixels. The comparison between all three methods was performed to select the one which gave the best results for measuring the ripeness of the banana.

8) Fruit Yield: Cao *et al.* [15] developed an Unmanned Aerial Vehicle (UAV) to capture multi-spectral images for

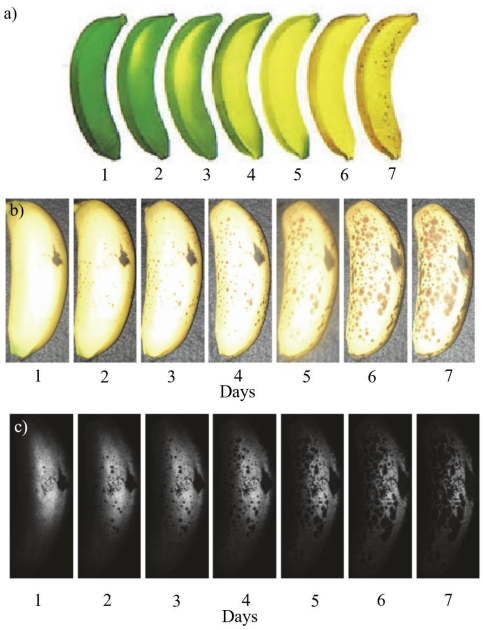


Fig. 7. (a) Seven Phases of the Ripening Process of the Banana. (b) Example of the Images used during the First Week which shows the Colour of the Sample. (c) Example of the Images used during the First Week that shows Gray Scale Images from a Specified Spectral Region for the Same Sample [14].

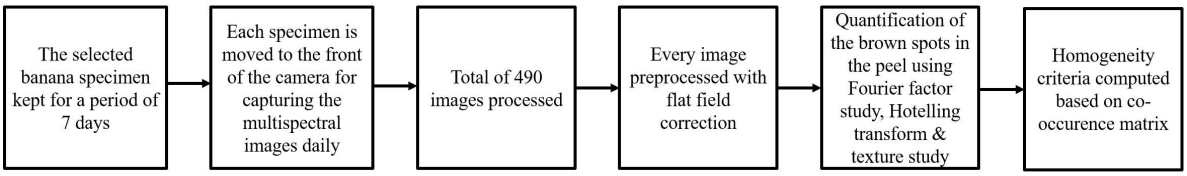


Fig. 8. Block Diagram to Estimate Ripeness of Banana Fruit.

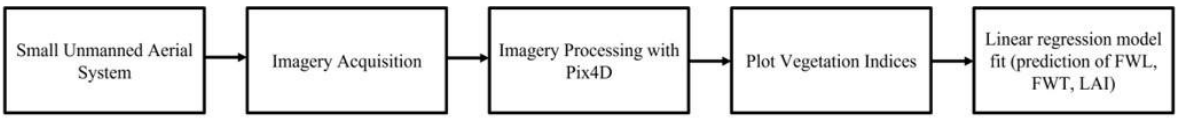


Fig. 9. Block Diagram of Fruit Yield of Sugar Beet Crops.

monitoring growth indicators of sugar beet crops. For this purpose, a wide dynamic range vegetation index (WDRVI) is used. The values achieved for the determination coefficients (R^2) for leaf area index (LAI), Fresh Weight of Leaves (FWL), and Fresh Weight of Roots (FWR) models of the sugar beet were created using the WDRVI were 0.957, 0.950, and 0.963, respectively. The accuracy of growth indicators of sugar beet can be improved from 1.05 % to 5.07 % using the WDRVI index model. The relation between the ground biomass of sugar beet and the growth indicator helps in finding the accuracy of the growth indicator monitoring. The increase in the ground biomass of sugar beet reduces beet growth indicator, improving the accuracy of growth monitoring by using WDRVI saturation. There is huge biomass in sugar beet and thus, this technique proves to be useful. Wittstruck *et al.* [16] used UAV-based image data to detect Hokkaido pumpkins and estimate their yield with high accuracy.

9) Block Diagram to Find Sugar Beet Growth: A small UAV was used for aerial remote sensing as shown in Fig. 9. This comprised of MicaSense Red Edge-M multi-spectral sensor and a passive light optical sensor. The images of sugar beet were captured during the crucial growth period which consists of seedlings, leaf plumps growth, root growth, and accumulation period of sugar. The pre-processing of multi-spectral data was performed to remove noise, distortion in the sensor, removal of background information, and radial correction. 20 points were considered in the test area. In addition to this, a light sensor, GPS, inertial measurement unit was utilized to get the exact spatial data required. WGS84 was used as a geographic coordinate system. By adding the weight coefficient to NDVI, four WDRVI indices were calculated. This was done to evaluate the LAI, FWL, and FWR of the study plots. The sensitivity evaluation was performed based on the five vegetation indices and three growth indicators when NDVI $>$ 0.8. From the results, it was observed that the WDRVI index is more sensitive than NVDI. It was also seen that the

correlation coefficient between the beet growth index and the WDRVI index is higher than that of NDVI. The models were developed using the data from the years 2018 and 2019. 70% of the data was used to develop the model and 25% was used for verifying the model. RMSE, Relative Root Mean Square Error (RRMSE), R^2 were calculated. The results show that the WDRVI1 index with a weighted coefficient of 0.05 accurately evaluated the LAI, FWL, and FWR. WDRVI1 evaluated the FWL in the early growth phase in the best way.

C. Classification

Naeem *et al.* [17] applied multi-spectral imaging and texture feature extraction along with a Multi-Layer Perceptron (MLP) classifier to categorize medicinal plant leaves using five spectral bands with a range of 460 nm to 1560 nm. Six medicinal plant leaves were considered. Fig. 10 shows the grayscale image of the six medicinal plant leaves. The classification accuracy achieved for tulsi was 99.10%, for peppermint 99.80%, for Bael 98.40%, for lemon balm 99.90% , for catnip 98.40%, and for Stevia 99.20% using the multi-layer perceptron classifier. The MLP method achieved a high accuracy of 99.01% as compared to the other methods.

1) Block Diagram of Medicinal Plant Leaves Classification: A computer vision-based experimental setup was used for collecting multi-spectral and digital images as shown in Fig. 11(a). The images were transferred into the grey level format with a resolution of 800 × 800 pixels by cropping the leaf region exactly. A Sobel filter was used for edge/line detection of seed intensity. If the seed intensity threshold value is greater than 6, then mark the region as a region of observation. In this work, five regions of observation were drawn on each image and fused features were extracted from the data set. The chi-square feature selection technique reduces the feature vector space and selects the fourteen best features for medicinal plant leaves classification. Five machine learning classifiers namely MLP, LogitBoost (LB), Bagging (B), Random Forest (RF), and Simple Logistics (SL) were employed on the medicinal plant leaves database. Abdollahnejad *et al.* [18] proposed a method for tree species classification and health condition assessment with the help of Unmanned Aircraft System (UAS) multi-spectral imaging for the mixed broad leaf-Conifer forest. They developed a method to recognize healthy, unhealthy, and dead trees affected by bark beetle infection. The results show an overall accuracy (OA) = 81.18% and values of Kappa = 0.70 which prove the capability of their method to classify tree species. The health status condition of tree species shows the value of an OA = 84.71% and Kappa = 0.66. The SVM method proved as a good classifier of tree species. The fusion of vegetation indices (VI) and texture analysis (TA) layers also results in increased OA by 4.24%. Their method can be used to check huge areas affected by biological intrusion factors for mapping and detection very fast. Other applications of their proposed method are to estimate habitat conditions and tree inventory at low costs.

2) Block Diagram of Tree Species Classification and Health Status: UAS bi-temporal aerial imagery was used which incorporates a high-resolution 5-band spectrum as shown in Fig. 11(b). In the pre-processing step, the MicaSense RedEdge-M multi-spectral equipment was used to calibrate the reflectance panel and to convert raw pixel values into reflectance. This was performed to correctly describe the target consisting of the object of interest from the image pixel values. In the absence of calibration, the data collected will not be correctly compared for change identification. Using Structure from Motion (SfM) imaging photogrammetric database was generated. The Agisoft Metashape is used so that the values of image pixels correctly identify the region of interest by adjusting the atmospheric and light conditions For the canopy spectral study and texture study, DTMs, DSMs, and orthophotos were given as input with a resolution of 0.05m. For the spectral study, 9 VI and 13 TA variables were used. After this bi-temporal data was combined with texture data. Then supervised machine learning technique was employed to catch the reflectance values within a buffer area of 2m radius from each treetop. Using the same buffer area effect of crown shape and vague pixels from the overlapped crown areas were removed. The SVM algorithm delivered good results and could achieve the correct and reliable classification of tree species.

3) Fruit or Flower Detection: Vasas *et al.* [19] developed a technique for edge detection in bees in which multi-spectral images of flowers show the importance of utilizing long-wavelength sensitive receptors. With the help of multi-spectral image dataset of flowering plants, SNR ratios of long-wavelength (L) receptor reply were four times higher than the short (S) and medium (M) wavelength receptors depending on the specific conditions. The band-pass filters included the full wavelength range of bee vision which had transmission peaks at 340, 400, 460, 520, 580, 640, and 700 nm. As shown in Fig. 17 honeybees and bumblebees collect visual data with the help of 3 types of photoreceptors. However, only a single type of receptor is used that responds to longer wavelengths that are used for movement and edge detection. The results show peak sensitivity at 544 nm which corresponds to green for long-wavelength receptors which give the most steady signals in response to the natural objects.

4) Block Diagram to Detect Long Wavelength Receptor Responses of Bees: The database consisting of 53 images was scanned and processed in the ImageJ software as shown in Fig. 18. Java plugin color correction was used to adjust the white balance. The normalized cross-correlation (NCC) method was used to separate the images of the same flower. Different parts of the flowers such as petals, centers, and leaves were manually selected for comparing the reflectance spectra. A total of 52 images were captured. 1000 pixels were randomly selected from the different parts of the flower using which the standard deviation, mean, and SNR were calculated. To understand the difference between the vision of humans and bees, false-colored images were added to the database. The black and white images were tinted to the hue set by the filter wavelength with the help of ImageJ. The resultant images were combined with the RGB color images. The combination of RGB layers produces an original color in human vision. However, for bees, the UV, Blue (B), and yellow (Y) layers are tinted with B, Green (G), and Red (R) to produce false-color images. **Calculating Receptor Responses:** The Standard deviation (SD) and mean was calculated for the quantum catches (P) and relative receptor responses (E) which are functions of receptor peak sensitivities. Lastly, the effects of reflectance functions, illumination spectrum, background reflectance spectrum, the shape of the sensitivity of the receptor were evaluated. These results were combined

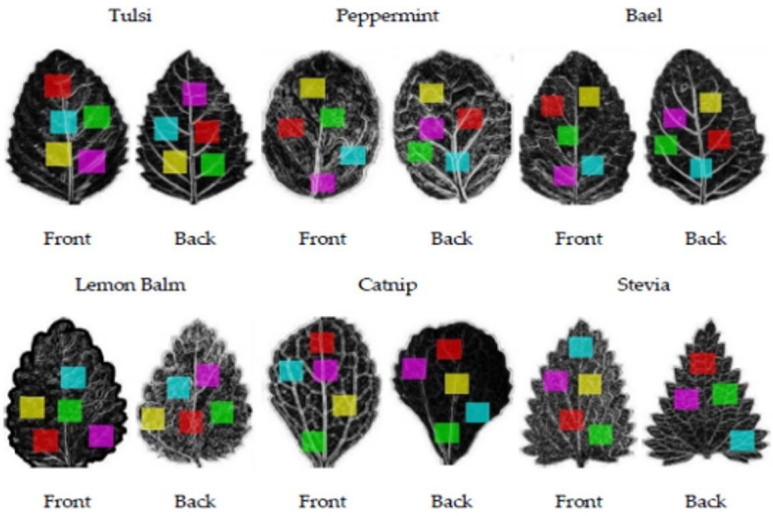


Fig. 10. Images used for Medicinal Plant Leaves Classification. The Original Images are Transformed into Grayscale with the ROI that Allows the Classification of these Leaves into Tulsi, Peppermint, Bael, Lemon Balm, Catnip and Stevia. It also shows Five Colorful Regions of Observation [17].

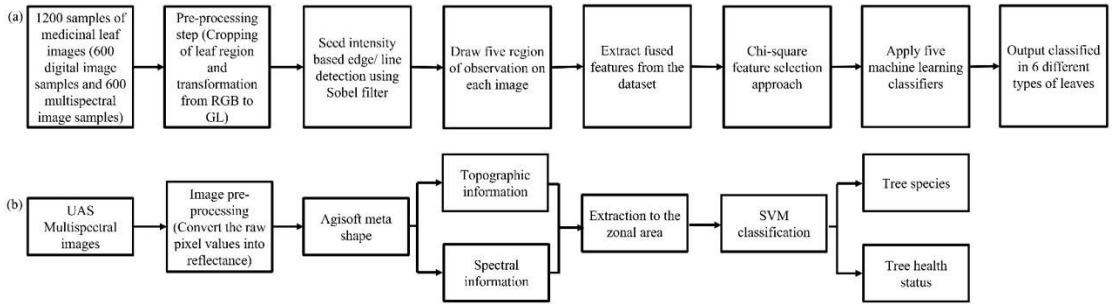


Fig. 11. Different Classification Strategies using Multi-Spectral Imaging on the Plant Data. (a) Block Diagram for Medicinal Plant Leaves Classification. (b) Block Diagram of Disease Detection in Tree Species to Recognize Healthy, Unhealthy and Dead Trees Affected by Bark Beetle Infection.

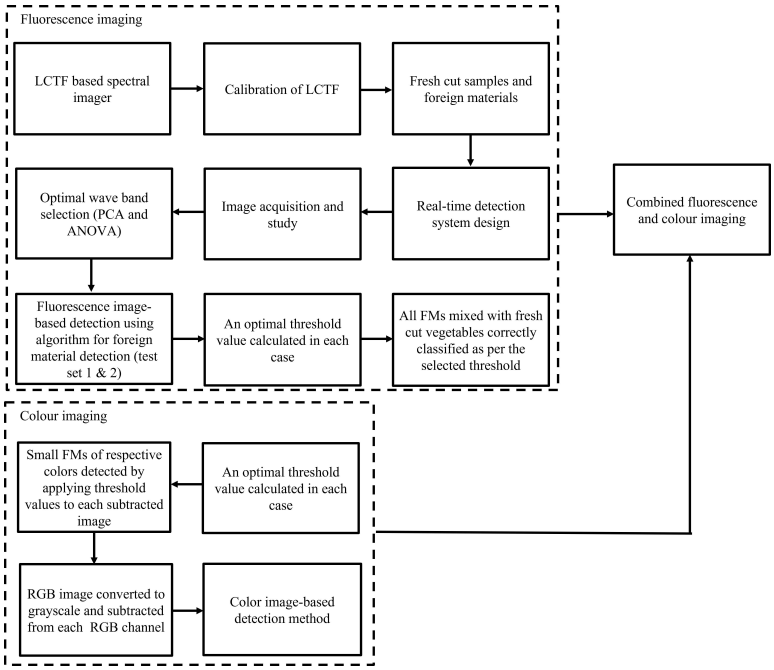


Fig. 12. Block Diagram for Detection of Foreign Material in Cabbage and Green Onion Samples using Reflectance and Fluorescence Images.

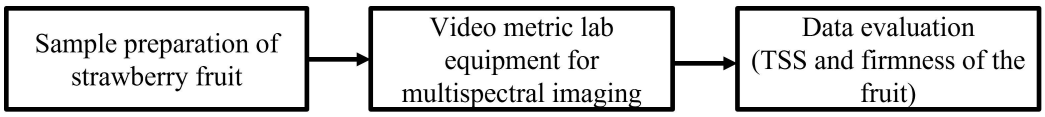


Fig. 13. Block Diagram of Internal Fruit Parameter in Strawberry Fruit (Firmness and TSS).

and given as input to the null model. To get a more clear picture of the bee's vision, heat maps were included which show the receptor responses of the long, medium, and short-wavelength for each multi-spectral image. A real-time system was developed by Lohumi *et al.* [20] for the identification of foreign materials (FMs) mixed with fresh-cut vegetables, with the help of fluorescence and color imaging. The setup included a multi-spectral fluorescence imager merged with a Liquid Crystal Tunable Filter (LCTF) to capture desired band images sequentially of fresh-cut vegetables. Fig. 13 shows the color images of FM with cabbage sample. The detection accuracy (average) of FMs in cabbage and green onion sample was considered. The average of the total detection accuracy for 4 repetitions was calculated to get the detection accuracy, which outperformed 95%. A processing unit for vegetables that were fresh-cut was installed in an industrial environment to further test the real-time detection system. The performance of the developed real-time system was almost the same in the industrial condition. The system can scan an area of a maximum of 24 × 24 cm^2 in a time duration of approximately 1.5 s.

5) Block Diagram to Identify FM in Fresh Cut Vegetables: As shown in Fig. 12 the LCTF based spectral imager was employed to focus a UV-A beam light onto the target area. A built-in white light was used to provide an alternative for the changeover. Thus, both reflectance and fluorescence images could be captured with this setup. **Fresh cut samples and FMs:** In this work, cabbage and green onion samples were used in response to the fluorescence and change in color as shown in Fig. 19. The FMs that were added to the freshly cut vegetables are plastic pieces, peel residues, small metal pieces, and various kinds of woods pieces, toothpicks, etc. **Real time detection system design:** White LED illumination was used to capture the color images of the samples on the conveyor unit. Fluorescence and color imaging units were placed in such a manner that they did not interfere with each other. The movement of the conveyor belt was controlled using a motion controller board and a computer unit and synchronized with the sensing unit. This was done to capture the fluorescence images of the samples placed on the conveyor unit. This was followed by, capturing the color images of the same area when the sample moved one step ahead, which corresponds to approx. 24 cm, of the Field of View (FOV) of the color camera. **Classification using optimal waveband selection:** The PCA and ANOVA methods were used to reduce the volume of data and select the optimal wavelength. In this study, differentiation of vegetables and FMs on the conveyor belt as well as the differentiation of the vegetable samples from the FMs was required. Thus, the best wavelengths were chosen to visually inspect the fluorescence spectra of the conveyor belt, fresh-cut vegetables, and FMs. The maximum peak intensity centered at a wavelength of 465 nm was chosen from the fluorescence spectra of the conveyor belt. The fluorescence peak intensity value was either 465 nm or 615 nm for different FMs. The fluorescent FMs have been recognized using these 2 band wavelengths. But the 465 nm waveband created a problem for fluorescent FMs as the images were partly covered by the fluorescent signals of fresh-cut vegetables. To solve this problem, a 435 nm wavelength image was used where a little variation is observed between the fluorescent intensity of fresh-cut vegetables and the FMs. Later these two band images were averaged (435 nm and 465 nm) and given a threshold to check the existence of fluorescent FMs. **Flourescence image-based detection:** Two tests were performed with an algorithm to detect the FMs. A threshold value was set in each case of the two tests by examining the smallest number of resulting false positive and false negative pixels. Based on the threshold, the classification of the FMs mixed with fresh-cut vegetables was done accurately with minimum error. **Color imaging-based detection:** Initially, the RGB image was transformed into a grayscale image, and the image was subtracted from each RGB channel. The threshold values which were set were applied to each subtracted image to detect the small FMs of corresponding colors. For the smallest number of resulting false positive and negative pixels, the best threshold value was selected in each case. It was observed that the accuracy of the color imaging-based detection method is less than that of fluorescence imaging. **Combination of Color and fluorescence imaging:** The combination of fluorescence and color imaging gives better accuracy. This technique gives much higher accuracy as compared to that of the single imaging technique. By counting the mutually independent detected FMs from the RGB-based and fluorescence-based classification images, the combined accuracy was calculated. In this work, combined detection accuracy (average) for 4 repetitions of the test exceeded 95%.

D. Disease Detection

Zhang *et al.* [22] used multi-spectral imaging for the detection of rice sheath blight (ShB) disease with the help of an unmanned aerial system. The results indicate that between ground-measured Normalized Difference Vegetation Index (NDVIs) and image-extracted NDVI's there was a strong correlation. The determination coefficient (R^2) is 0.907 and the RMSE is 0.0854. Whereas the results indicate that between image extracted NDVIs and disease severity there is a good correlation with a value of (R^2) = 0.627 and the RMSE = 0.0852. NDVIs calculated from the images achieved an accuracy of 63% which could measure various levels of the disease in the plots. The study included original RGB and hue, lightness, and saturation (HLS) transformation images of 67 field areas.

1) Block Diagram for Rice ShB Disease Detection: **Unmanned aerial system (UAS):** To capture the imagery data from the field plot, a UAV along with a higher resolution multi-spectral camera was employed (refer Fig. **??**). The Micasense

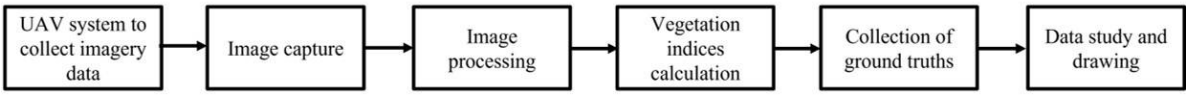


Fig. 14. A Generalized Approach for Rice Sheath Blight (ShB) Disease Detection using Multi-Spectral Imaging to Estimate Severity on a Rated Scale (0 to ±9).

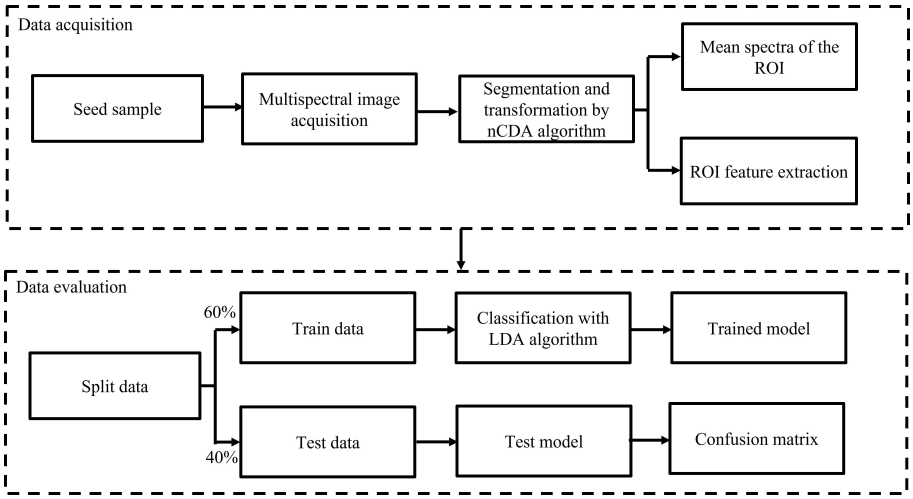


Fig. 15. Block Diagram to Differentiate between Fungi-Infested and Non-Infested Black Oat Seeds using Spectral Information.

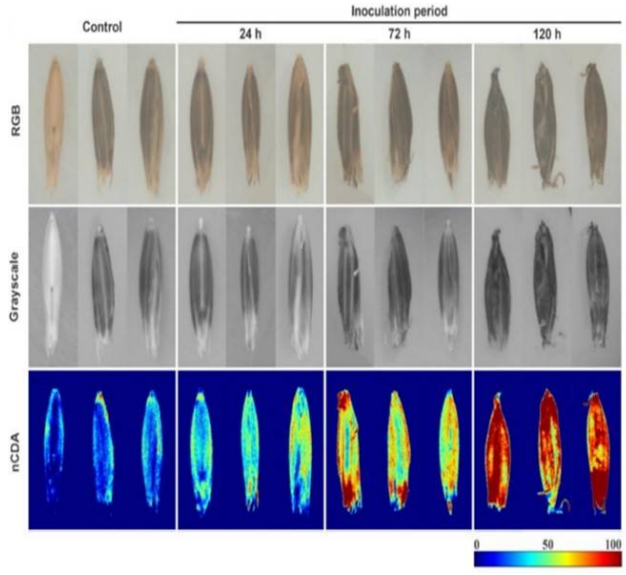


Fig. 16. Images of Black Oat Seeds: Raw Images and Corresponding Grayscale and nCDA at Wavelength 365 nm for Seeds Free from Fungus and Seeds Exposed to D. Avenae Fungi for 24, 72, and 120 Hours. The Blue Color Defines Healthy Tissues, Intermediate Contamination is Defined by Green and Yellow Colors and the Red Color Defines Large Contamination in Fungi in the Images Transformed by the nCDA Algorithm [21].

RedEdgeTM can capture 12-bit raw images in 5 narrow bands. This camera was used to calculate the reflectance of the plant which contains data about stress in crops more accurately than the 3 channel camera. **Image capture:** To cover all 67 field plots the camera was directed at the nadir at two heights, 27 meters above the ground and 5.5. meter to cover 4 plots for every image during the flight. **Image pre-processing:** To acquire color traits from the captured images and conversion of these images into various color spaces, ENVI software was employed. In this work, digital images of various severity levels of ShB were converted to lightness, hue, and saturation. After this, the mean values of HLS were taken. **VI calculation:** 5 types of VI's - NDVI, Ration Vegetation Index (RVI), Difference Vegetation Index (DVI), Normalized Difference Water Index (NDWI), and Red Edge (RE) were measured from the captured multi-spectral images. Change maps of various disease severity stages were produced, so that the image information could find the indications and the ShB disease severity to scale it. **Ground truth information gathering:** The ground NDVI readings were taken with the help of the GreenSeeker handheld crop sensor. Various measurements were recorded in every field plot so that the accuracy of NDVI values increased which signified the severity of the ShB disease. In this work, 134 average NDVI readings were taken from the areas of 67 plots that were inoculated and un-inoculated with pathogens. After this ShB severity was rated on a scale of 0±9 concerning disease symptoms. **Data evaluation and drawing:** For UAV image processing Pix4D mapper was used. ArcGIS 9.1 was used for mapping and geospatial study. For statistical evaluation PASW Statistic 18, software was used. RMSE and determination coefficient were used to find the accuracy of the correlation model. Françca *et al.* [21] developed a model for identifying D.Avenae fungi in black oat seeds using multi-spectral images. The study was performed using color and texture parameters from seeds incubated for 120 hours. Fig. 15 shows the raw image and relative grayscale and nCDA of black oat seeds. Results show the high performance of the model with an accuracy of 0.86. This indicates that the multi-spectral imaging method was capable of recognizing D.Avenae fungi in black oat seeds. The accuracy achieved was 0.86 for the color and texture feature which was satisfactory. The accuracy achieved for black oat seeds inoculated for 24, 72, and 120 hours was 0.78, 0.83, and 1.00 respectively. This shows that multi-spectral imaging can be an adequate method that will assure that the black oat seeds do not contain any diseases or fungi.

2) Block Diagram to Detect D.Avenae Fungi in Black Oat Seeds: **Data Acquisition:** With the help of the Videometer-Lab4 instrument, 19 multispectral images of 200 non-infested seeds as well as 200 infested seeds were acquired (Fig. 16 c)). These seeds were infested with D. Avenae fungi for an inoculation interval of 24, 72, and 120 hours. 19 high-resolution images were acquired at a time in 5 seconds. VideometerLab4 software version was used to analyze the data. The nCDA algorithm was employed to transform the images. This was performed so that there exists a maximum separation between the classes and minimum distance within a specific class. For every seed, the ROI was recognized and a mask was created to segment the seeds from the surrounding. The blob database was used to collect the seeds and from each seed, 36 variables were obtained. This also involved color components such as hue and saturation. The earlier acquired 19 multi-spectral images were then used to obtain the color features. Various color description models were employed to transform the multi-spectral data into color data. By calculating the individual spectral bands, the texture features were obtained. The 19 spectral wavelengths are in the range of 365 nm to 970 nm. The mean reflectance of each seed lying within the range of the spectral bands was extracted with the help of the MultiColorMean feature. A trimmed mean of the transformed pixel values was calculated using RegionMSI (mean). Similarly, the percentage of the blob region that had a transformation value higher than the threshold was calculated using RegionMSIthresh. **Data Evaluation:** LDA method was used to develop two models that could differentiate between infested and non-infested seeds from the spectral information. The reflectance data obtained for each seed was used to create the first model. Similarly, color and texture features were used to create the second model. The data was split into two sets with 60% of each sample used for training the model and the remaining 40% of each sample for testing the model. The result of the model was evaluated using the kappa coefficient and accuracy. The VideometerLab4 and R software were used for statistical evaluation.

III. Results

The results are summarised with the help of tables [Tables I to III] which concludes all the eight objectives of multi-spectral imaging of this manuscript in terms of their database size, number of color spectrums, accuracy, and other factors. Table I compares the various methods for varietal identification and quality analysis of fruits and vegetables. Shreshta *et al.* [8] has achieved the highest accuracy of 96% for the identification of varieties of 298 tomato seeds. Liu *et al.* classified 250 rice seeds samples based on their varieties. The accuracy achieved was 94%. For the quality analysis of pomegranate fruits, Khodabakhshian *et al.* [10] considered 100 samples and achieved the highest accuracy of 97% among the reported literature. Clemmensen *et al.* [9] had used the 747 number of samples for fruit quality analysis, which was the largest dataset used. A comparison of reported literature for leaf and tree species classification along with disease detection is shown in Table 2. The method proposed by Naeem *et al.* [17] for lemon balm leaf classification achieved a classification accuracy of 99.9 % on a dataset of 600 images. Among the methods proposed for disease detection, Franca-Silva *et al.* [21] had achieved the highest accuracy of 100% for 400 samples by using 19 color spectrums. They had used the largest number of samples and the largest number of color spectrums. Table 3 compares reported literature in the areas of fruit/flower detection, internal fruit parameters, fruit ripeness, and fruit yield. For detection of foreign material in cabbage and green onion samples, Lohumi *et al.* [20] achieved an accuracy of greater than 95 %. For extracting quality parameters for 210 strawberry fruits, Liu *et al.* [1] has achieved 100 % accuracy by using 19 different wavelengths. For classifying the fruit ripeness of 490 banana samples, Santoyo *et al.* [14] achieved the highest accuracy of 84 %. For predicting the fruit yield of sugar beet crops, Cao *et al.* [15] achieved the highest accuracy of 96.3 % for 15 field plots.

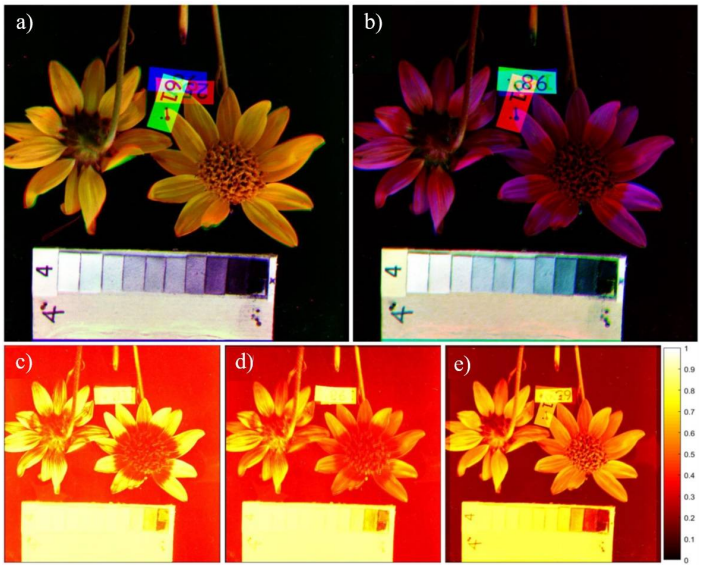


Fig. 17. A Series of Monochromatic Filters having Different Peak Transmissions are used to Capture Black-and-White Images. Further, a False-Color (FC) Image is Produced from these Photos to Correct the White Balance from the Images. These FC Images are Later Tinted with Black and White Images to Get a Proper Hue. (a) The 3 Layers i.e. RGB are Combined for Human Vision. (b) Transformation of Yellow, Blue and UV into Red, Green and Blue is Performed for the Bees. The Receptor Response is Calculated for the (c) Short (d) Medium (e) and Long-Wavelength Receptors which Explained the Electrical Signals that the Photoreceptors in Bees Generate in response to the Image [19].

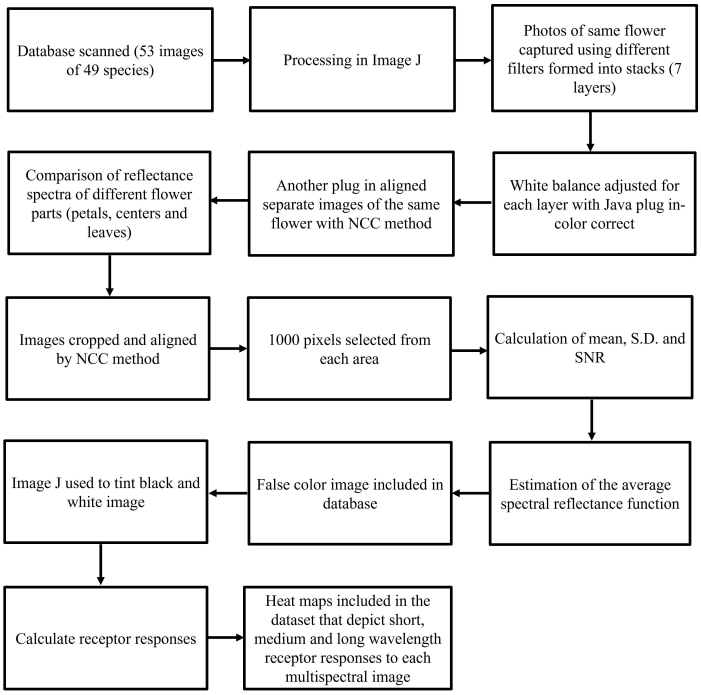


Fig. 18. Block Diagram of Fruit/Flower Detection using Long-Wavelength Sensitive Receptor for Bees.

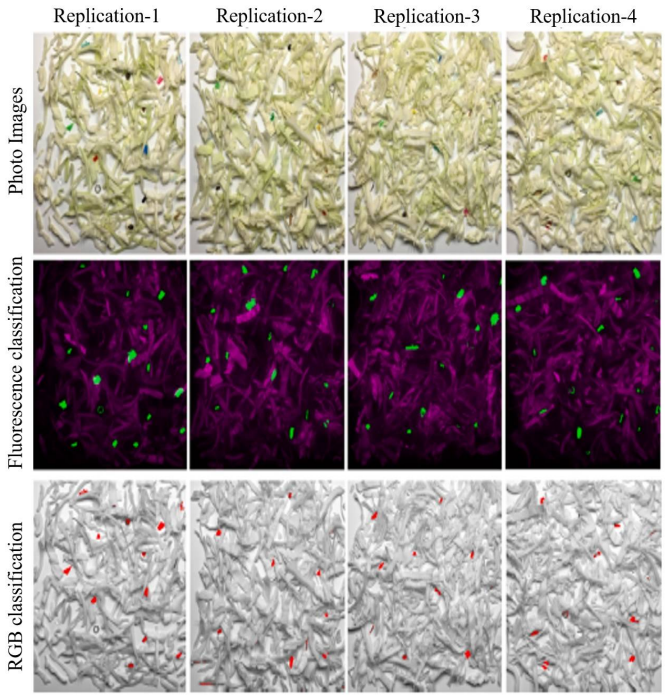


Fig. 19. Colour Images of Cabbage Samples Mixed with Foreign Materials (FM) are shown in the Top Row. Classification Images based on Fluorescence are shown in the Middle Row. Classification Images based on RGB are shown in the Bottom Row [20].

TABLE I. VARIETAL IDENTIFICATION AND FRUIT QUALITY

Sr. No.	Method	Database size	Number of colour spectrums used	Accuracy (in %)
Varietal Identification				
1.	Liu et al. [13]	250 seeds	19	94
2.	Shrestha et al. [8]	298 seeds	19	96
Fruit Quality				
1.	Khodabakhshian et al. [10]	100 samples	4	97
2.	Clemmensen et al.[9]	747 samples	20	-
3.	Lianou et al.[12]	245 samples	18	91.7

IV. RESEARCH GAPS OF REPORTED LITERATURE

For varietal identification of rice seeds, PLS-DA and PCA-BPNN models achieved low accuracies. PLS-DA and PCA-BPNN, even when tremendously used in chemometrics, does not give good results as in LS-SVM, at least for this case [1]. The spectroscopy method calculates the total amount of transmitted or reflected light from a certain part of a sample and does not contain wavelength-specific data [23]. This limitation was solved by using the hyperspectral imaging system which spatially receives the wavelength responses (pixel) of images of the fruit [10]. For each wavelength and each piece of vegetable, a t-test was conducted to check for significant differences on a 5% level in several percentiles of the light reflectance. The results show significant changes from days 2 to 4 in the reflectance spectrum for both the celeriac and carrots, significant changes continuing until day 14 were found [9]. A threshold value was set as TVC ≤ 2 to assess the quality of the vanilla cream samples in the dairy industry to indicate a fresh sample. In earlier methods, finding the threshold limit was a costly, effort-intensive, and lengthy process. Thus, multispectral imaging provides a less expensive, fast, and automatic method that helps in the quick judgment process for quality managers [24]. To determine the quality features in the strawberry fruit, the limitation was that the NIR spectrometers can only identify a small part of the fruit not the entire fruit at certain times. The limitation of hyperspectral imaging is that problems are faced while processing the data that make it difficult for commercial real-time applications [25]. This problem is solved using multispectral imaging which enables a fast, nondestructive examination of the interior and exterior features of different types of vegetables and fruits [13]. The earlier methods that involve optical sensing, spectroscopy, and imaging are more complex in their algorithms or experimental setups, which are not useful in real-time applications. Other than this, earlier techniques never incorporated multispectral imaging in their work. So, this work focuses on multispectral imaging which helps in finding, recording, and quantifying the banana ripening process [14]. For observation of sugar beet growth indicators, WDRVI was used. Earlier studies used satellite platforms that are confined to both height and orbit and cannot manage the temporal, spatial, or spectral resolution essential for monitoring the growth [26]. The new UAV tech-

TABLE II. CLASSIFICATION AND DISEASE DETECTION

Sr. No.	Method	Database size	Number of colour spectrums used	Accuracy
Classification				
1.	Naeem et al.[17]	600 images	5	99.9
2.	Abdollahnejad et al.[18]	297 images	5	84.71
Disease Detection				
1.	Zhang et al.[22]	67 field plots	5	63
2.	Franca-Silva et al.[21]	400 samples	19	100

TABLE III. FRUIT/FLOWER DETECTION, INTERNAL FRUIT PARAMETER, FRUIT RIPENESS/MATURITY AND FRUIT YIELD.

Sr. No.	Method	Database size	No of colour spectrums used	Accuracy (in %)
Fruit/ flower detection				
1.	Vasas et al [19]	53 samples	7	-
2.	Lohumi et al.[20]	-	2	> 95
Internal fruit parameter				
1.	Liu et al.[1]	210 fruits	19	100
Fruit ripeness/ maturity				
1.	Santoyo et al.[14]	490 images	2	84
Fruit Yield				
1.	Cao et al. [15]	15 plots	5	96.3

nology can manage low-cost, crop growth observation at high temporal, spatial, and spectral resolution. The RGB images acquired by the UAV evaluated the canopy estimated area of a tree and provided the best option. The LAI and leaf cluster biomass found from NDVI is used to monitor growth, but this is not possible for medium and high bio-masses. A WDRVI developed uses NVDI NIR band reflectance data, which causes lowering of the weights in such bands which belong to high and medium biomass cases [27]. This results in increased linearity and reduces saturation, which makes correct growth monitoring easier. An enhanced WDRVI is used to evaluate the early-phase growth of sugar beet [15]. In the classification of medicinal plant leaves, the feature selection (FS) process is the most crucial part of the ML-based classification. The goal of this research is to achieve better accuracy in less time. It is observed that without feature selection, the MLP classifier takes a lot of time (4.83 Seconds) to identify medicinal plants due to a large number of features. But when selected features are used, higher accuracy is obtained in less time. The PCA method is an unsupervised approach and does not provide good results on labeled data, as in this work the medicinal plant leaf varieties data set is labeled. This limitation in PCA is resolved by ML-based supervised feature selection techniques: chi-square feature evaluator (ranked search) method [28]. This method selects the best features from the huge features vector space (FVS). Compared to PCA, the current method can extract the sub-database with the best features for this large database [17]. Tree species classification studies are generally focused on rural and urban forests instead of managed forests. They are also limited to only spectral study and usually incorporate RGB sensors. Work on plant nurseries and low vegetation which has low tree volume and low levels of design complexity are useful, but forest conditions create problems for detection. SVM method was used with a combination of textural and spectral data. Later statistical methods for classifying tree species and identification of dead and unhealthy trees harmed by bark beetles are investigated [18]. The multispectral images allow us to model the visual response that the bees' photoreceptors give in response to the flower image. This, in turn, will help to differentiate the function of the different wavelength channels in color processing [19]. The earlier methods for detecting foreign objects in vegetables were metal detection, X-ray inspection, and color imaging techniques are not satisfactory. Also, the machine vision technology with the color camera is not useful for visibly opaque plastic FMs, and cannot detect the FMs that are similar in color to fresh-cut vegetables. The limitations of the earlier work were solved by using fluorescence imaging in this work [20]. The rice Sheath Blight (ShB) disease can be identified by using UAS, RGB, and multispectral imagery data. The color features computed using the multispectral images and the RGB images could hardly identify the variations in covering caused due to the disease [29]. Also, the information available is not enough to classify various levels of sheath blight, as the wavelength range is small and large bands related to the RGB camera are used [22]. For detection of D. Avenue fungi in black oat seeds, currently, the detection is done by examining the dry seed visually. This method is hard, slow, and requires experts. Advanced sensors combined with object recognition by computers that can automatically operate will result in a fast examination of seed health status by extraction of wavelength, texture features, and color [30]. Multispectral imaging combines optical spectroscopy and computer vision, which results in producing spatial and wavelength-specific data on different fungi species [21].

V. FUTURE WORK

A. Varietal Identification

For varietal identification of tomato, further studies will focus on the use of spectral data related to VIS-NIR obtained from multi-spectral imaging to discriminate the rice seeds of different varieties. This information will be related to practical organic aspects of tomato and other crops [8].

B. Fruit Quality

For quality evaluation of pomegranate fruit, future research will concentrate on the execution of the developed multi-spectral imaging by carrying out the real-time test [10]. For wok-fried vegetables of carrot and celeriac samples, the sensory quality of the wok-fried celeriac and carrots is assessed using the Quality Index Method (QIM). The scores of all features are then added that results in an overall sensory score. This is called Quality Index. QIM indicates scores of 0 for fresh products and larger scores in increasing order for spoiled products. This plan is observed by images and complete information of all factors. The sensory study requires that the samples are reheated before assessment, as otherwise, it is not possible to assess taste and smell [9]. For vanilla cream, microbiological quality classification the Unsupervised Online Feature Selection (UOS) algorithm could manage a classification error that was quite low in the test database. The threshold limit which was set for the "fresh" samples of vanilla cream, was a crucial factor for product delivery to the market by the manufacturer. This example gives a good outlook for the application of multi-spectral imaging in combination with the UOS algorithm in the quality management of the dairy industry [12].

1) Internal Fruit Parameter: To find quality features and ripeness phase in strawberry fruit, PLS, SVM, and BPNN models were employed. All the models' used visible areas of the spectrum. But, if the NIR spectra are also incorporated then the results could be marginally improved. The multi-spectral imaging system can reduce the image retrieval as well as operating time in contrast to the hyper-spectral imaging system that allows online automated quality observation systems [13].

2) Fruit Ripeness/Maturity: For the ripening process in bananas, future work can be online applications to measure maturity level in other fruits [14].

3) Fruit Yield: For observation of sugar beet growth indicators, the future work can be for growth tracking of potato, radish, and other underground crops to a large extent for reaping benefits [15].

C. Classification

For the classification of medicinal plant leaves, the reported study was limited to six medicinal plant leaves while there are millions of types of medicinal plant/herbs in the world. This is a pixel-based method and in the future, an object-based method can be used. In the future, this proposed method can be used on other medicinal plant leaves. Further, the results can be improved using hyper-spectral and 3D digital image data sets [17]. For tree species classification and health status assessment, identifying physical stress in earlier stages or finding small wavelength changes can be done with high reliability and accuracy by using hyper-spectral sensors [31]. This was not possible in multi-spectral sensors as it does not operate in that wavelength range. Future work can be on the application of thermal sensors which will enhance the classification and health status of different tree species. The combination of multi-spectral sensors and UAS can decrease the operating costs of tracking forest areas in the small to medium range [18].

1) Fruit/Flower Detection: For multi-spectral images of flowers, receptors with long wavelengths for edge detection in bees were used. The results provide acceptable information for selecting the long-wavelength channel over the shorter-wavelength channels for clarifying a visual scene. This can initiate the ways environment and prior knowledge can build more resulting neural processes which can be the future work [19]. For real-time detection of foreign material mixed with fresh-cut vegetables, two types of fresh-cut vegetables were tested. For future work, this method can be experimented on different vegetables by picking the best fluorescence wavelengths(band) and using a suitable algorithm for processing images of fresh-cut vegetables [20].

D. Disease Detection

For detection of rice sheath blight, using a UAV can help in the cultivation of rice cultivars with resistance to Sheath Blight disease. For future work, a new UAV system developed can be used for site-specific accurate fungicide application technique to control ShB disease in rice [22]. For the detection of D. Avenae in black oat seeds, multi-spectral imaging can be used in the future for separating seed-carrying fungi, affected by other elements, but are firmly related to physical and chemical differences caused by fungi [21].

VI. Conclusion

In this manuscript, we have multi-spectral imaging which is the emerging technology for the grading of fruits and vegetables. Quality assessment of vegetables and fruits is the need of the hour for the food industry. To meet consumer demand and profit, the technologies developed should provide more concise, fast, and accurate results. Some of these technologies have their merits and demerits. We have reviewed different ways in which multi-spectral images could be used for fruit and vegetable assessment like varietal identification which includes identification of rice seeds and tomato, fruit quality which includes wok-fried vegetables(carrots and celeriac), quality of vanilla cream, etc. Such images can also be used classification of medicinal plant leaves and tree species, rice sheath blight disease detection, identifying pathogens in black oat seeds, etc. Multi-spectral imaging also enables the detection of foreign material in cabbage, to find how the honey bees and bumblebees respond to images of flowers for long, short, and medium wavelength receptors. Internal fruit parameters like firmness and TSS in strawberries can also be extracted. Other applications of multi-spectral imaging include the estimation of fruit ripeness and fruit yield. The data obtained by multi-spectral imaging can enable the evaluation of the classification of fruits and vegetables. The study of multi-spectral imaging for fruit and vegetable will be beneficial to disease prevention, irrigation, and yield improvement. Furthermore, multi-spectral information can be considered the basis for evaluating various characteristics and parameters of fruit, vegetable, and tree species. It would also promote research on fruit, vegetable, and trees species.

Declarations

A. Funding Information

No funding was involved in the present work.

B. Conflicts of Interest

Authors S. Gaikwad and S. Tidke declare that there has been no conflict of interest.

C. Code Availability

Not applicable.

D. Authors' Contributions

Conceptualization was done by Shilpa Gaikwad (SG) and Sonali Tidke (ST). All the literature reading and data gathering were performed by SG. The formal analysis was performed by SG. Manuscript writing and original draft preparation was done by SG. Review and editing was done by ST. Visualization work was carried out by SG and ST.

E. Ethics Approval

All authors consciously assure that the manuscript fulfills the following statements: 1) This material is the authors' own original work, which has not been previously published elsewhere. 2) The paper is not currently being considered for publication elsewhere. 3) The paper reflects the authors' own research and analysis in a truthful and complete manner. 4) The paper properly credits the meaningful contributions of co-authors and co-researchers. 5) The results are appropriately placed in the context of prior and existing research.

F. Involvement of Human Participant and Animals

All the necessary permissions were obtained from the Institute Ethical Committee and concerned authorities to run our algorithms on patient data.

G. Information about Informed Consent

Informed consent was obtained from participants whose data was used to do analysis.

H. Consent for Publication

Authors have taken all the necessary consents for publication wherever required.

REFERENCES

[1] W. Liu, C. Liu, F. Ma, X. Lu, J. Yang, and L. Zheng, "Online variety discrimination of rice seeds using multispectral imaging and chemometric methods," *Journal of Applied Spectroscopy*, vol. 82, no. 6, pp. 993–999, 2016.

[2] T. Bourlai and B. Cukic, "Multi-spectral face recognition: Identification of people in difficult environments," in *2012 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2012, pp. 196–201.

[3] M. Moroni, A. Mei, A. Leonardi, E. Lupo, and F. L. Marca, "Pet and pvc separation with hyperspectral imagery," *Sensors*, vol. 15, no. 1, pp. 2205–2227, 2015.

[4] H.-W. Chen, J.-H. Lee, B.-Y. Lin, S. Chen, and S.-T. Wu, "Liquid crystal display and organic light-emitting diode display: present status and future perspectives," *Light: Science & Applications*, vol. 7, no. 3, pp. 17 168–17 168, 2018.

[5] J. D. Vincent, S. Hodges, J. Vampola, M. Stegall, and G. Pierce, *Fundamentals of Infrared and Visible Detector Operation and Testing*. John Wiley and Sons, 2015.

[6] I. Makki, R. Younes, C. Francis, T. Bianchi, and M. Zucchetti, "A survey of landmine detection using hyperspectral imaging," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 124, pp. 40–53, 2017.

[7] B. Zhang, W. Huang, J. Li, C. Zhao, S. Fan, J. Wu, and C. Liu, "Principles, developments and applications of computer vision for external quality inspection of fruits and vegetables: A review," *Food Research International*, vol. 62, pp. 326–343, 2014.

[8] S. Shrestha, L. C. Deleuran, M. H. Olesen, and R. Gislum, "Use of multispectral imaging in varietal identification of tomato," *Sensors*, vol. 15, no. 2, pp. 4496–4512, 2015.

[9] L. H. Clemmensen, B. S. Dissing, G. Hyldig, and H. Løje, "Multispectral imaging of wok-fried vegetables," *Journal of Imaging Science and Technology*, vol. 56, no. 2, pp. 20 404–1, 2012.

[10] R. Khodabakhshian, B. Emadi, M. Khojastehpour, M. R. Golzarian, and A. Sazgarnia, "Development of a multispectral imaging system for online quality assessment of pomegranate fruit," *International Journal of Food Properties*, vol. 20, no. 1, pp. 107–118, 2017.

[11] A. Bhargava and A. Bansal, "Grading of variety of bi and mono-colored applest," in *Soft Computing and Signal Processing*. Springer, 2022, pp. 375–382.

[12] A. Lianou, A. Mencattini, A. Catini, C. Di Natale, G.-J. E. Nychas, E. Martinelli, and E. Z. Panagou, "Online feature selection for robust classification of the microbiological quality of traditional vanilla cream by means of multispectral imaging," *Sensors*, vol. 19, no. 19, p. 4071, 2019.

[13] C. Liu, W. Liu, X. Lu, F. Ma, W. Chen, J. Yang, and L. Zheng, "Application of multispectral imaging to determine quality attributes and ripeness stage in strawberry fruit," *PloS one*, vol. 9, no. 2, p. e87818, 2014.

[14] M. Santoyo-Mora, A. Sancen-Plaza, A. Espinosa-Calderon, A. I. Barranco-Gutierrez, and J. Prado-Olivarez, "Nondestructive quantification of the ripening process in banana (musa aab simmonds) using multispectral imaging," *Journal of Sensors*, vol. 2019, 2019.

[15] Y. Cao, G. L. Li, Y. K. Luo, Q. Pan, and S. Y. Zhang, "Monitoring of sugar beet growth indicators using wide-dynamic-range vegetation index (wdrvi) derived from uav multispectral images," *Computers and Electronics in Agriculture*, vol. 171, p. 105331, 2020.

[16] L. Wittstruck, I. Kühling, D. Trautz, M. Kohlbrecher, and T. Jarmer, "Uav-based rgb imagery for hokkaido pumpkin (cucurbita max.) detection and yield estimation," *Sensors*, vol. 21, no. 1, p. 118, 2021.

[17] S. Naeem, A. Ali, C. Chesneau, M. H. Tahir, F. Jamal, R. A. K. Sherwani, and M. Ul Hassan, "The classification of medicinal plant leaves based on multispectral and texture feature using machine learning approach," *Agronomy*, vol. 11, no. 2, p. 263, 2021.

[18] A. Abdollahnejad and D. Panagiotidis, "Tree species classification and health status assessment for a mixed broadleaf-conifer forest with uas multispectral imaging," *Remote Sensing*, vol. 12, no. 22, p. 3722, 2020.

[19] V. Vasas, D. Hanley, P. G. Kevan, and L. Chittka, "Multispectral images of flowers reveal the adaptive significance of using long-wavelength-sensitive receptors for edge detection in bees," *Journal of Comparative Physiology A*, vol. 203, no. 4, pp. 301–311, 2017.

[20] S. Lohumi, B.-K. Cho, and S. Hong, "Lctf-based multispectral fluorescence imaging: System development and potential for real-time foreign object detection in fresh-cut vegetable processing," *Computers and Electronics in Agriculture*, vol. 180, p. 105912, 2021.

[21] F. França-Silva, C. H. Q. Rego, F. G. Gomes-Junior, M. H. D. d. Moraes, A. D. d. Medeiros, and C. B. d. Silva, "Detection of drechslera avenae (eidam) sharif [helminthosporium avenae (eidam)] in black oat seeds (avena strigosa schreb) using multispectral imaging," *Sensors*, vol. 20, no. 12, p. 3343, 2020.

[22] D. Zhang, X. Zhou, J. Zhang, Y. Lan, C. Xu, and D. Liang, "Detection of rice sheath blight using an unmanned aerial system with high-resolution color and multispectral imaging," *PloS one*, vol. 13, no. 5, p. e0187470, 2018.

[23] A. Ghita, P. Matousek, and N. Stone, "High sensitivity non-invasive detection of calcifications deep inside biological tissue using transmission raman spectroscopy," *Journal of biophotonics*, vol. 11, no. 1, p. e201600260, 2018.

[24] G. ElMasry, N. Mandour, S. Al-Rejaie, E. Belin, and D. Rousseau, "Recent applications of multispectral imaging in seed phenotyping and quality monitoring—an overview," *Sensors*, vol. 19, no. 5, p. 1090, 2019.

[25] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, and J. J. Sousa, "Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry," *Remote Sensing*, vol. 9, no. 11, p. 1110, 2017.

[26] W. H. Maes and K. Steppe, "Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture," *Trends in plant science*, vol. 24, no. 2, pp. 152–164, 2019.

[27] L. Wang, Q. Chang, F. Li, L. Yan, Y. Huang, Q. Wang, and L. Luo, "Effects of growth stage development on paddy rice leaf area index prediction models," *Remote Sensing*, vol. 11, no. 3, p. 361, 2019.

[28] S. Picek, A. Heuser, A. Jovic, and L. Batina, "A systematic evaluation of profiling through focused feature selection," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 12, pp. 2802–2815, 2019.

[29] S. Thomas, M. T. Kuska, D. Bohnenkamp, A. Brugger, E. Alisaac, M. Wahabzada, J. Behmann, and A.-K. Mahlein, "Benefits of hyperspectral imaging for plant disease detection and plant protection: a technical perspective," *Journal of Plant Diseases and Protection*, vol. 125, no. 1, pp. 5–20, 2018.

[30] S. Cubero, N. Aleixos, E. Moltó, J. Gómez-Sanchis, and J. Blasco, "Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables," *Food and bioprocess technology*, vol. 4, no. 4, pp. 487–504, 2011.

[31] A. Lowe, N. Harrison, and A. P. French, "Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress," *Plant methods*, vol. 13, no. 1, pp. 1–12, 2017.

Detecting Malware Families and Subfamilies using Machine Learning Algorithms: An Empirical Study

Esraa Odat and Batool Alazzam
Faculty of Computer and Information Technology
Jordan University of Science and Technology
Irbid, Jordan 22110

Qussai M. Yaseen*
College of Engineering and Information Technology,
Ajman University, UAE
Faculty of Computer and Information Technology
Jordan University of Science and Technology, Irbid, Jordan 22110

***Abstract*—Machine learning algorithms have proved their effectiveness in detecting malware. This paper conducts an empirical study to demonstrate the effectiveness of selected machine learning algorithms in detecting and classifying Android malware using permissions features. The used dataset consists of 9000 different malicious applications from the CIC-Maldroid2020, CIC-Maldroid2017 and CIC-InvesAndMal2019 datasets collected by the Canadian Institute for Cybersecurity. Meta-Multiclass and Random Forest ensemble classifiers are used based on different machine learning classifiers to overcome the imbalance in the data classes. Moreover, a genetic attribute selection technique and SMOTE are used to classify Ransomware sub-families to handle the small size of the dataset and underfitting problem. The results show that optimization and ensemble approaches are successful in treating dataset issues, with 95% accuracy in classifying big malware families and 80% in Ransomware subfamilies.**

***Keywords*—Malware classification; machine learning; SMOT; information security**

I. Introduction

Malware is a malicious software that aims at affects the confidentiality, integrity or availability of data and systems without users consent to attain the harmful intent of the attacker [1] [2]. Malware applications are classified into many classes according to their behaviour and properties such as adware, worms, viruses, rootkits, trojan horse, backdoor, spyware, logic bombs, adware, and ransomware. Systems resources are attacked to affect the assets for the purposes of getting financial benefits, for stealing private information or using the computing resources to attack other victims [3] [4] [5].

The usage of smartphone devices are growing immensely, which provides attackers a powerful mean to access users private information. According to google [6], there were 2 billion Android devices until November 2017, which means that Android operating system has 71.15% of the Mobile Operating System Market Share Worldwide [7] [8]. The wide spread of Android devices has increased Android attacks three times for the past two years. Therefore, there is a significant need to find ways to detect and classify malware families.

Fig. 1 shows the mobile malicious installation packages for Android that were discovered by Kaspersky Company only between 2017 to 2020, which shows an increase in the discovered malware from 2019 to 2020 by more than 2,100,000 malicious packages [9].

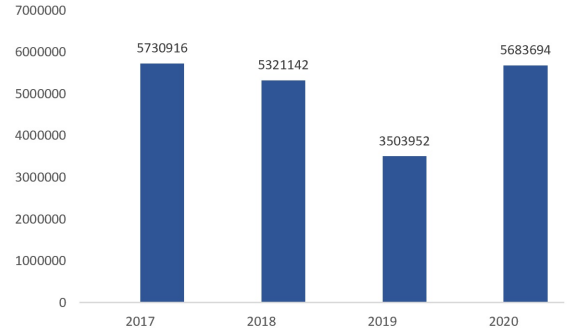


Fig. 1. Mobile Malicious Installation Packages for Android in 2017 through 2020

Android malware apps can be classified using static analysis method, dynamic analysis method, and Hybrid approach. In Static analysis method, the static features such as static APIs and permissions are used to classify APKs into malware or benign applications. Meanwhile, dynamic analysis approach uses dynamic features such as dynamic APIs, memory usage, CPU usage,Network outgoing traffic, etc. to classify malware and benign applications. Extracting dynamic features of malware is performed using an isolated environment, called sandbox, such as cuckoo sandbox [10]. Fig. 2 shows the environment of dynamic analysis of android applicatoins and extracting dynamic features. The hybrid approach uses a combination of static features and dynamic features to classify Android APKs.

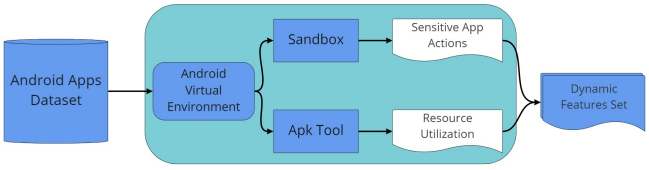


Fig. 2. Dynamic Analysis of Android Apps

The aforementioned approaches may be used with hyperparameter tuning to perform correctly. Similarly, Other approaches are used to deal with the underfitting problem, such as genetic algorithms [11]. The Genetic Algorithm (GA) is a search-driven optimization technique based on genetic and natural selection principles. It is often used to find almost the optimal solutions for complex problems [6]. Furthermore,

Malware analysis using machine learning may face dataset oversampling problems. For example, synthetic Minority Oversampling Technique (SMOTE) [12], which is an oversampling technique, is used to generate samples for the minority class. It cares about the feature space to generate new samples and helps of interpolation between positive samples that belong to each other. SMOTE selects the first positive class randomly then KNN's for the chosen positive sample is gained [7].

This paper implements and compare the performance of several machine learning algorithms in classifying a large dataset of APKs into malware and benign applications, classifying malware applications into its main classes, and classifying ransomware applications into its subfamilies. The used datasets are CIC-Maldroid2020, CIC-Maldroid2017 and CIC-InvesAndMal2019, which consists of 9000 Android APK samples were collected by the Canadian Institute for Cybersecurity.The measures used are f-score, accuracy, TPR and FPR. The contributions of this paper are summarized as follows:

- It classifies the dataset into malware or benign application using various machine learning algorithms.
- It classifies the malware into main classes: SMS malware, Ransomware, Banking malware, Scareware, Adware, and Premium SMS malware.
- It classifies the Ransomware family as subfamilies depending on extracted permissions as static features.
- It employs optimization techniques to overcome the problems dataset: ensemble algorithms, SMOTE, and genetic algorithm. The ensemble machine learning algorithms. The meta-Multiclass and Random Forest classifiers are used to solve unbalancing problems in classifying malware into its main classes to get accurate and unbiased performance measures. Furthermore,the paper applies the genetic algorithm and SMOTE used to improve the performance of classifying ransomware subfamilies because it is a small and imbalance dataset.

The rest of paper is organized as follows. The next section discusses some related work. Section 3 demonstrates the used methods and materials. Section 4 demonstrates and discusses the results. Finally, Section 5 concludes the work.

II. Related Work

The use of machine learning algorithms in detecting and classifying malware have been studied widely. However, due to the continuous growing of malware types, the change in their features, and the skills of attackers, machine learning algorithms and the features they use need to be optimized an modified continuously. This section discusses some related work in this field.

Zhiwu et al. [13] proposed a classification method based on static features, which are bytecode features, assembler code features, and PE features. The authors used eight machine learning classifier models, which are Gaussian Naïve Bayes, Random Forest, Decision Tree, SVM linear, SVM, KNN, and Ada-boosting. They used a dataset from VirusShare. In the comparison among the different algorithms, the cost, the needs, and convenience were considered. In addition, they used Kaspersky scan engine results. According to the experiments, the Random Forest model achieved the highest result of the F1 score of 93.56% .

Fang et al. [14] proposed an approach based on extracting control-flow graphs (CFG) and data-flow graphs (DFG) during the instruction level. Then, they encoded the graphs into matrices and used them to build the family classification model using deep learning. The family classification model considered the horizontal combination of CFG and DFG as features to achieve the best performance. The malware dataset used in the experiments were collected from Marvin, Drebin, VirusShare and Contagio-Dump. The results showed that the horizontal combination of CFG and DFG performed better than CDGDroid.

Arslan et al. [15] proposed a method for Android malware families classification using Dalvik Executable (DEX) file section features. The proposed approach converted DEX files to Red/Green/Blue images and plain text. Next, from these images, texture of the image, the color, and text were extracted as features. The results showed that the proposed method achieved a precision of 96%.

Gandotra et al. [16] used the manifest file to extract the permissions as features of android applications. The source code was used to verify whether a permissions is requested by the application or not. The considered the assumption that malicious apps are those that request several different permissions without using them. The proposed approach extracted all permissions requested by an application and stored them in a database. Then, the use of permissions was verified. Next, for each app, the proposed approach computed the number of suspicious values. These values were used for classification. The results showed that the SVs for malicious apps and benign apps were in the ranges [4-95,858] and [21-55,967] respectively. To classify the apps, the approach used Naive Bayes and Logistic Regression. The authors claimed that the accuracy was 91.95%.

Şahın et al. [17] proposed an approach to detect zero-day malware based on the integration of static and dynamic analysis. The authors validated the proposed approach using a real-world dataset of malicious samples. The performance of the proposed approach was measured before and after features selection to show the effectiveness of the proposed relevant feature selection method in improving the model construction time as well as the accuracy. The proposed approach used the entropy to compute the purity of each feature and select the relevant features. The results showed that the accuracy of all classifiers using the integrated features set is very good. However, the naive Bayes classifier did not achieve good results. Furthermore, the results showed that the approach of features selection improved the model building time and the accuracy.

Udayakumar [18] suggested a new permission-weight approach. The approach applied the algorithms of KNN and Naive Bays to build the model and used RF as a weighting matrix to assign weights for permissions based on whether a permission is requested by the app (benign and malicious apps). The results showed that by adding the weighted approach, the KNN algorithm improved 2% and the NB algorithm improved 7%.

Milosevic et al. [19] proposed an approach that help in understand the types of malware, and machine learning algorithms, such as Decision Trees, Multilayer Perceptron and Multi SVM, can be used to detect malware. The proposed approach used the debug Size as a feature with a score of 0.26. Using SVM and neural networks for classification, the proposed approach got an of 90.2% and 98% at training approach, and neural networks get 99% at the testing approach.

Alzubaidi [20] applied an approach based on deep learning algorithms to detect android malware apps families. He claimed that he achieved an accuracy of 99%. Meanwhile, Ashit [21] applied an enhanced Birch algorithm to find the malware and modified executables of Windows and Android operating system.

III. Methods and Materials

The method proposed in this paper is divided into two parts: classifying main malware-families and classifying Ransomware subfamilies. The classification process is based on using permissions as features to show the ability of machine learning classifiers to distinguish between the different malware families. Furthermore, it aims at helping security officers to develop more reliable systems against different types of malware, and to detect malicious applications by using only the permissions and minimum resources.

A. Features Extraction

The "apktool.jar", which is a reverse engineering tool, was used to extract the permissions of each APK. As a result, 140 permissions were extracted. Next, a python code was developed and used to represent the APKs and the requested permissions by each APK in a vector space file.

Fig. 3 shows the operations of the extracting and selecting permissions features. The figure shows that the APK tool was used to analyze the Android APK file in order to extract the Manifest.xml and classes.dex files. Next, the Manifest parser was used to extract the permission features, from which the used features were selected form the classification process.

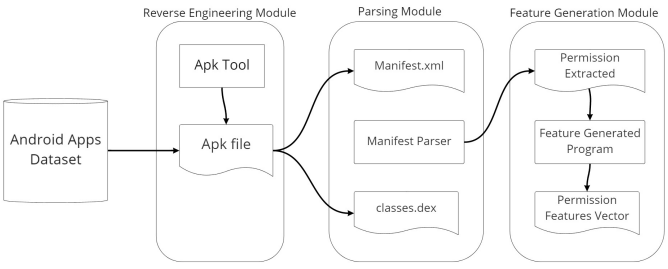


Fig. 3. Extracting and Selecting Permissions Features

B. Dataset

Two datasets have been used in this work. The first dataset is CIC-Maldroid2020 [23] collected from Canadian Institute for Cybersecurity. The dataset consists of 8750 Android samples as APK files, while the second dataset consists of 250 ransomware APKs collected from CIC-InvesAndMal2019 [22] and CIC-Maldroid2017 [24]. The dataset was divided into three subsets. The first dataset consists of 1422 Adware, 2506 Banking, and 4821 SMS malware. The second dataset consists of 250 Ransomware subfamilies.

C. Methodology

The work has two main parts. The first one is the feature extraction of APKs permissions as well as feature selection. The second one is the data processing, where main malware families were classified using ensemble machine learning algorithm; Random Forest, Decision Tree and meta-Multiclass classifiers. Meanwhile, Naïve Bayes, KNN, Decision Tree, and Logistic Regression were used to classify Ransomware family as sub-families.

IV. Results and Analysis

A. Main Malware Family Classification

The classification of malware APKs into main classes were performed using Meta-Multiclass, which is a classifier for handling multi-class datasets with 2-class classifiers. Meta-Multiclass has various capabilities including error correcting output codes to increase the accuracy. Using this method, when the base classifier cannot handle instance weights because they are not uniform, the data is re-sampled with replacement based on the weights before being processed by the base classifier.

Table I shows the performance matrices for the main malware families classification. The result shows that the Meta-Multiclass classifier achieved high performance by classifying malware with 94% accuracy and low error rate of about 0.029.

Fig. 4 shows classification error for meta-Multiclass classifier using KNN as base classifier. The x-axis and y-axis represent the main malware family names, where 1, 2 and 3 labels denotes Adware, SMS and Banking families respectively, star shapes represent correct predicted points and square shapes represent incorrect predicted points. The results show that the highest prediction power of the classifier is for Adware data samples with 97% true positive rate.

TABLE I. Meta-Multiclass Classifier

Malware Family	TPR	FPR	F -SCORE	ROC
Adware	0.968	0.018	0.966	0.988
SMS	0.872	0.022	0.910	0.954
Banking	0.985	0.047	0.947	0.979
AVG	0.942	0.029	0.941	0.974

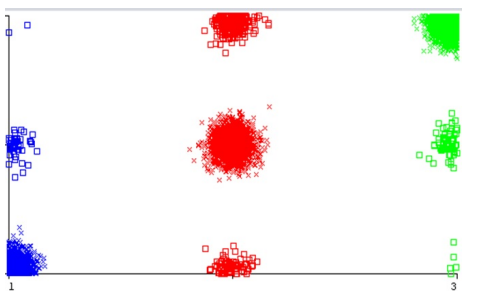


Fig. 4. Classification Error of Meta-Multiclass Classifier using KNN base Classifier

Fig. 5 shows the classification error for Random Forest classifier. The x-axis and y-axis represent main malware family names, where 1, 2 and 3 labels denotes Adware, SMS and Banking families respectively, star shapes represent correct predicted points and square shapes represent incorrect predicted points. To proof the meta-Multiclass findings, random forest ensemble classifier on decision tree as a base classifier was applied. According to the results, the two used ensemble methods showed similar performance.

Using Meta-Multiclass classifier based on KNN as base classifier and Random Forest classifier, the accuracy of ensemble models are 95%. Furthermore, the results show that the ensemble classifiers handle the data unbalancing problem and classify main malware families correctly. The values of TPR, F-score, accuracy, and ROC show the high performance and low FPR value of about 0.027 as shown in Table II.

TABLE II. META-MULTICLASS CLASSIFIER BASED ON KNN AS BASE CLASSIFIER AND RANDOM FOREST CLASSIFIER

Malware Family	TPR	FPR	F -SCORE	ROC
Adware	0.977	0.018	0.970	0.993
SMS	0.875	0.019	0.915	0.964
Banking	0.983	0.045	0.949	0.981
AVG	0.945	0.027	0.945	0.979

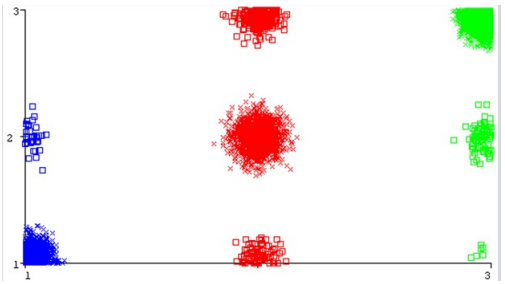


Fig. 5. Classification Error of Random Forest Classifier

B. Ransomware Subfamilies Classification

The dataset used in this experiment was imbalanced. Therefore, the challenge was to apply Machine learning algorithms to measure the effectiveness of ML different models in distinguishing ransomware subfamilies and treating the data problems.

1) Ransomware Subfamilies Classification using Original Dataset: To evaluate the effectiveness of machine learning algorithms in detecting Ransomware subfamilies, the Decision tree classifier was applied on the original dataset.

Fig. 6 shows the classification error for Decision Tree classifier on original dataset, were the x-axis and y-axis represent Ransomware subfamily names, star shapes represent correct predicted points and square shapes represent incorrect predicted points.

Table III shows that using Decision Tree classifier, the accuracy are 62%. That is, the results show that machine learning algorithms based on permissions requested by applications on original dataset does not effectively classify Ransomware subfamilies.

TABLE III. DECISION TREE CLASSIFIER

Sub-family	TPR	FPR	F-score	ROC
Wannalocker	0	0.02	0	0.88
Svpeng	0.8	0.13	0.6	0.89
Simplocker	0.5	0.03	0.52	0.79
RansomBO	0.6	0.06	0.5	0.94
PornDroid	0.6	0.13	0.55	0.87
Pletor	0.5	0.04	0.5	0.88
LockerPin	0.8	0	0.9	0.91
Koler	0.5	0	0.66	0.91
Jisut	0.9	0.01	0.92	0.98
Charger	0.8	0.01	0.84	0.99
Average	0.6	0.06	0.61	0.9

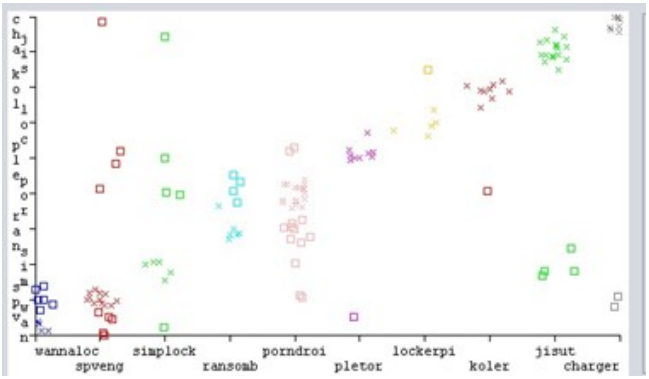


Fig. 6. Classification Error of Decision Tree on Original Data

2) Ransomware Subfamilies Classification using Optimization Algorithms: To address data underfitting and imbalance issues, we used Synthetic Minority Oversampling Technique (SMOT), which uses the KNN algorithm to generate new observations to eliminate the imbalance. In addition, we used the genetic algorithm as hyperparameter tuning or parameter tuning. Furthermore, the paper used the ensemble learning algorithm to increase the model chance learning and produce a good prediction.

Fig. 7 shows the classification error for Random Forest classifier with optimization algorithms, where x-axis and y-axis are Ransomware subfamily names, star shapes represent correct predicted points and square shapes represent incorrect predicted points. The imbalance and underfitting problems were almost solved after applying SMOT and the genetic algorithms with the random forest classifier. We noticed that the accuracy improved up to 80% from the previous results, which was 60%, with low false positive rate with 0.028 as shown in Table IV.

TABLE IV. RANDOM FOREST CLASSIFIER

Sub-family	TPR	FPR	F-score	ROC
Wannalocker	0.4	0.05	0.39	0.94
Svpeng	0.7	0.05	0.69	0.97
Simplocker	0.9	0	0.94	0.96
RansomBO	0.7	0.07	0.54	0.95
PornDroid	0.7	0.03	0.78	0.97
Pletor	1	0	1	1
LockerPin	0.8	0	0.9	0.99
Koler	0.9	0	0.94	0.99
Jisut	1	0.02	0.95	0.99
Charger	1	0	1	1
Average	0.8	0.03	0.8	0.97

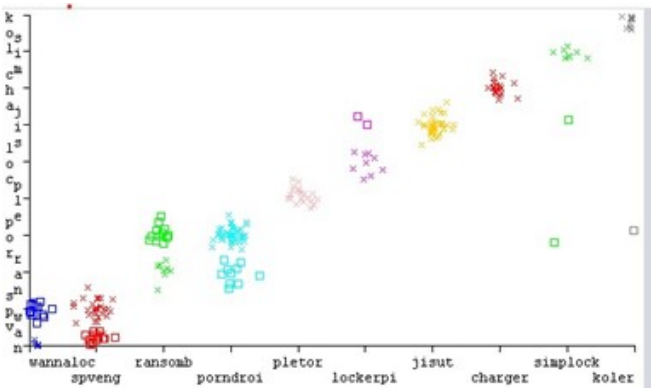


Fig. 7. Classification Error of Random Forest with Optimization Algorithms

V. Conclusion

Android threats and attacks are rapidly increasing as Android devices and the number of users increasing around the world. Therefore, the attacks on android operating systems and users private information has increased. This paper has discussed this issue and used an optimized machine learning approach based on permissions as features to detect malware. The paper has used a large dataset to detect malware and classify detected malware into main malware families, which are SMS malware, Banking malware, Adware. In addition, the optimized approach was tuned to classify ransomware dataset into its subfamilies. The optimized approach used ensemble classifiers such as meta-Multiclass classifier with KNN as base classifier and Random forest to classify main malware families and handling the unbalanced dataset. In addition, the optimized approach used Random Forest and Decision Tree classifiers to classify ransomware subfamilies. The results have shown that ensemble classifiers perform very well in handling unbalanced data, detecting and classifying main malware families by achieving an accuracy of 95%. Furthermore, the results have shown that detecting and classifying Ransomware subfamilies using traditional machine learning algorithms has a poor performance with an accuracy of 62% for Decision Tree. However, the tuned approach by the oversampling technique has increased the accuracy to 81%.

References

[1] C. K. A. Moser en E. Kirda, “Limits of Static Analysis for Malware Detection”, in 23rd Annual Computer Security Applications Conference, 2007, pp 421–430.

[2] E. Kirda, “Dynamic Analysis of Malicious Code”, Journal in Computer Virology, vol 2, pp 67–77, 2006.

[3] “A comparison of static, dynamic, and hybrid analysis for malware detection”, Journal in Computer Virology, vol 13, pp 1–24, 2017.

[4] E. M. Dovom, A. Azmoodeh, A. Dehghantanha, D. E. Newton, R. M. Parizi, en H. Karimipour, “Fuzzy pattern tree for edge malware detection and categorization in iot”, Journal Systems Architecture, vol 97, pp 1–7, 2019.

[5] M. Ficco en F. Palmieri, “Leaf: An open-source cybersecurity training platform for realistic edge-iot scenarios”, Journal Systems Architecture, vol 97, pp 107—129, 2019.

[6] B. Popper, “https://www.theverge.com/2017/5/17/15654454/android-reaches-2-billion-monthly-active-users”, 2017 (accessed,2021).

[7] StatCounter, “Mobile Operating System Market Share Worldwide”, 2020 (accessed,2021).

[8] B. Sun, Q. Li, Y. Guo, Q. Wen, X. Lin, en W. Liu, “Malware family classification method based on static feature extraction”, in 2017 3rd IEEE International Conference on Computer and Communications (ICCC), 2017, pp 507–513.

[9] Kaspersky, “Malware Growth”, Available: https://securelist.com/mobile-malware-evolution-2020/101029/. (accessed,2021).

[10] “Cuckoo Sandbox”.Available: https://cuckoosandbox.org,(accessed,2021).

[11] Rodkaew Y., “A Genetic Algorithm as a Classifier”, in 18th International Conference on Electrical Engineering/Electronics , Computer, Telecommunications and Information Technology (ECTI-CON), 2021, pp 254–257.

[12] “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary”, Journal of Artificial Intelligence Research, vol 61, pp 863–905, 2018.

[13] X. U. Zhiwu, K. Ren, en F. Song, “Android malware family classification and characterization using CFG and DFG”, in 2019 International Symposium on Theoretical Aspects of Software Engineering (TASE), 2019, pp 49–56.

[14] Y. Fang, Y. Gao, F. Jing, en L. Zhang, “Android malware familial classification based on DEX file section features”, IEEE Access, vol 8, pp 10614–10627, 2020.

[15] R. S. Arslan, İ. A. Doğru, en N. Barişçi, “Permission-based malware detection system for android using machine learning techniques”, International journal of software engineering and knowledge engineering, vol 29, no 01, pp 43–61, 2019.

[16] E. Gandotra, D. Bansal, en S. Sofat, “Zero-day malware detection”, in 2016 Sixth international symposium on embedded computing and system design (ISED), 2016, pp 171–175.

[17] D. Ö. Şahın, O. E. Kural, S. Akleylek, en E. Kiliç, “New results on permission based static analysis for Android malware”, in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp 1–4.

[18] N. Udayakumar, V. J. Saglani, A. V. Cupta, en T. Subbulakshmi, “Malware classification using machine learning algorithms”, in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 2018, pp 1–9.

[19] N. Milosevic, A. Dehghantanha, en K.-K. R. Choo, “Machine learning aided Android malware classification”, Computers and Electrical Engineering, vol 61, pp 266–274, 2017.

[20] A. Alzubaidi, “Sustainable Android Malware Detection Scheme using Deep Learning Algorithm”, International Journal of Advanced Computer Science and Applications, vol 12, pp 860–867, 2021.

[21] A. Dutta, “Detection of Malware and Malicious Executables Using E-Birch Algorithm”, International Journal of Advanced Computer Science and Applications, vol 7, pp 124—126,

[22] Canadian Institute of Cybersecurity, “CIC-InvesAndMal2019”, Available: https://www.unb.ca/cic/datasets/invesandmal2019.html, 2019 (accessed,2022).

[23] Canadian Institute of Cybersecurity, “CIC-Maldroid2020”, Available: https://www.unb.ca/cic/datasets/maldroid-2020.html, 2020 (accessed,2022).

[24] Canadian Institute of Cybersecurity, “CIC-Maldroid2017”, Available: https://www.unb.ca/cic/datasets/andmal2017.html, 2017 (accessed, 2022)

Systematic Exploration and Classification of Useful Comments in Stack Overflow

Prasadhi Ranasinghe, Nipuni Chandimali, Chaman Wijesiriwardana
Faculty of Information Technology
University of Moratuwa
Katubedda, Sri Lanka

***Abstract*—Stack Overflow is a public platform for developers to share their knowledge on programming with an engaged community. Crowdsourced programming knowledge is not only generated through questions and answers but also through comments which are commonly known as developer discussions. Despite the availability of standard commenting guidelines on Stack Overflow, some users tend to post comments not adhering to those guidelines. This practice affects the quality of the developer discussion, thus adversely affecting the knowledge-sharing process. Literature reveals that analyzing the comments could facilitate the process of learning and knowledge sharing. Therefore, this study intends to extract and classify useful comments into three categories: request clarification, constructive criticism, and relevant information. In this study, the classification of useful comments was performed using the Support Vector Machine (SVM) algorithm with five different kernels. Feature engineering was conducted to identify the possibility of concatenating ten external features with textual features. During the feature evaluation, it was identified that only TF-IDF and N-grams scores help classify useful comments. The evaluation results confirm Radial Basis Function (RBF) kernel of the SVM classification algorithm performs best in classifying useful comments in Stack Overflow regardless of the usage of the optimal combinations of hyperparameters.**

***Keywords*—Stack overflow; useful comments; machine learning; SVM; classification**

I. Introduction

The software engineering community considers Stack Overflow as a learning site and a learning community for software developers and practitioners [1], [2]. Comments in Stack Overflow (SO) are temporary *"Post-It"* notes relevant to a particular question or an answer which has already been posted [3]. They clarify and enrich the content conveyed through questions and answers. Examining comments is particularly beneficial because they go beyond the questions and answers to facilitate the process of learning and knowledge construction. For example, the importance of analyzing Stack Overflow comments for the recommendation of source code fragments has been extensively discussed in the literature [4], [5], [6].

In Stack Overflow, there is a standardized method to add comments which are suggested in the standard commenting guidelines of Stack Overflow. *Request clarification*, *constructive criticism*, and *relevant but transient information* in the comments are being encouraged in commenting guidelines in Stack Overflow [7]. Comments related to *request clarification* contain requests for extra information for better understanding the post. *Constructive criticism* comments point out flaws, obsolescence, and coding errors thus encouraging the author to improve it. Comments related to *relevant but transient information* guide the users in retrieving more information that is relevant to a certain post in Stack Overflow [7]. Nevertheless, Stack Overflow does not recommend comments related to suggesting corrections, compliments, answering a question, criticisms, secondary discussions, and discussions of site policies and community behavior to be posted on the site[1]. However, it is observed that most of the developers tend to respond to posts with comments which are not adhering to Stack Overflow's guidelines on comments [8]. Hence, useful comments get ignored by the authors of relevant posts as well as the members of the community. According to previous studies, 27.5% of the comments which required an update from the author were ignored [9]. Therefore, comments in Stack Overflow should be studied in depth to identify whether users post comments by adhering to the commenting guidelines of Stack Overflow.

It is believed that systematic categorization of such comments could provide valuable insights to software practitioners when using Stack Overflow as a learning source. Thus, there is a need for data-driven solutions to retrieve useful comments and categorize them. However, this direction has not been thoroughly investigated in the literature [7], [10], [11]. Therefore, this study exploits machine learning and natural language processing to automatically classify the comments in Stack Overflow that follow the standard commenting guidelines based on three types: request clarification, constructive criticism, and relevant information. This paper expects to address the following research question (RQ):

RQ : How to correctly classify the useful comments in Stack Overflow into standard comment categories: request clarification, constructive criticism, and relevant information?

The remainder of this paper is organized as follows. Next, we present the related work of this study. Then the methodology of classifying useful comments in Stack Overflow is presented. Next, the results and evaluation are described. Finally, the conclusion of the study and the further work are being discussed.

II. Related Work

During the past years, researchers have involved themselves in various studies related to comments and crowd source knowledge in Stack Overflow. The author in [7] analyzed commenting activities in respect of timing, content and individuals who perform commenting focusing on the comments

[1] https://stackoverflow.com/help/privileges/comment

which were posted in answers in Stack Overflow. In their study, comment classification was done by a light-weight open coding process in which authors were involved in deriving a draft list of comment types. The time in which users take to post comments once the answer was posted was taken into consideration. They mention the need of a methodical and organized study related to the comments in Stack Overflow to better understand how comments are being used. Moreover, improving the current commenting system was stated as necessary since users post comments in unrecommended manners. Furthermore, they state the possibility of future research to leverage approaches from Machine Learning and NLP communities to automatically identify such comment categories [7].

In their research, Sengupta and Haythornthwaite performed a qualitative analysis for the purpose of introducing a coding schema which comprises nine comment categories through classification. The purpose of their study was the provision of insights into commenting in Stack Overflow. They mention the requirement of further research on comment usage [1]. Identification of inadequate comments in Wikipedia's talk page edits and classification of the same into different categories was performed by Sulke and Varude. In this analysis, SVM provided the best results out of the utilized classification algorithms [12]. Contextual tagging mechanism was utilized to classify posts in Stack Overflow in the exploration done by Chimalakonda et.al. SVM promised the highest accuracy of 78.5% in this approach. In this analysis, Limited number of posts were examined during the study. Furthermore, Some of the statistical distributions were not balanced and they were biased towards one certain topic [13].

Beyer et.al in their study automated the classification of Stack Overflow's posts into seven categories of questions. Manual Analysis of phrases was performed to find patterns and training classification models based on Machine Learning Algorithms was done in their analysis. Since manual analysis was performed there exists a possibility for this categorization of the posts to be biased [14]. Saif et.al performed online toxic comment classification by making use of three Artificial Neural Network Approaches and Logistic Regression [15]. Quantitative as well as a qualitative analysis related to the obsolete answers of Stack Overflow was conducted in an Empirical study related to the obsolete knowledge on Stack Overflow. The utilized heuristics based approach contained the accuracy of 75%. In this analysis, it was believed that Machine Learning could provide better results [7]. In a study related to the prediction of who will answer a specific question in Stack Overflow, a hybrid approach which amalgamates both knowledge from the question and the asker to retrieve more error-free candidate lists was considered as essential [16]. In a mining approach which suggests insightful comments in Stack Overflow, recommendation of source code comments was accomplished with the accuracy and precision of 80%. In this approach, Dataset which was considered for the empirical evaluation was limited. Furthermore, it was stated that the approach might be inadequate in recommending comments for the code segment from proprietary or legacy projects [4]. The value of comments revolves around many important aspects such as improving the source code, analyzing the code further, and facilitating code reuse [4] [17] [18] [19]. During the Exploration of the means of which comments have an effect on answer updates, comments and answer updates which involve code segments were only being taken into consideration. Moreover, it was mentioned that there exists a tendency for the comments to be mislabeled if the code element in the comment is not correctly identified by the system, which leads to false positives [9]. A Gold Standard for Emotion Annotation in Stack Overflow was introduced by Novielli et.al. In this exploration, manual annotation of Stack Overflow gold standard data set with emotion labels was performed. Identification of Emotions was based on clear guidelines of a conceptual framework which is based on theory. Moreover, Final gold labels were assigned through agreement of the majority of 3 coders [20].

Automatic comment generation approach which mines comments from Q&A sites includes code description mapping extraction, refinement of description, code clone detection, code clone pruning and selection of comments. Failure at identifying comments that comprise an incorrect description of the code segment was stated as a limitation in this study [21]. In the study that extracted candidate method documentation from discussions of Stack Overflow, JavaDoc descriptions were created and the mining of source code descriptions from developers' discussions was recognized as an aspect which needed more improvement when regarding its usability [22]. Wikipedia's comment dataset by Jigsaw was used in analysing any section of text and detecting distinct types of toxicity by Chakrabarty. In this analysis, TF-IDF with 6 headed Machine Learning promised the highest accuracy which is 98.98%. Chakrabarty states that the utilization of Grid search Algorithm can obtain more accurate results [23].

III. Methodology

Useful comments in Stack Overflow can be defined as the comments, which adheres to Stack Overflow's commenting guidelines. The facilitation of the process of learning and knowledge construction could be improved by analyzing useful comments. Classification is an approach for analyzing useful comments. This study aims at extracting useful comments and classifying them into standard useful comment categories based on their features. In this study, Data Extraction, Qualitative Analysis and Review, Feature Engineering and Preprocessing, Feature Extraction, Training and Evaluation of the classification model were carried out. Fig. 1 represents the high-level architectural diagram of exploration of useful comments in Stack Overflow.

A. Data Extraction

This step was necessary to obtain the comments data which is needed to perform the study. In this study comments posted in the Stack Exchange Data Explorer during the past 5 years (i.e., between 1st of January 2015 and 08th of November 2020 were taken into consideration).

Listing 1: Query utilized in obtaining comments data from Stack Exchange Data Explorer

```
SELECT Comments.Id, Comments.PostId, Comments.Score,
       Comments.Text, Comments.CreationDate,
       Comments.UserId, Comments.ContentLicense,
       Posts.Tags
FROM   Comments
```

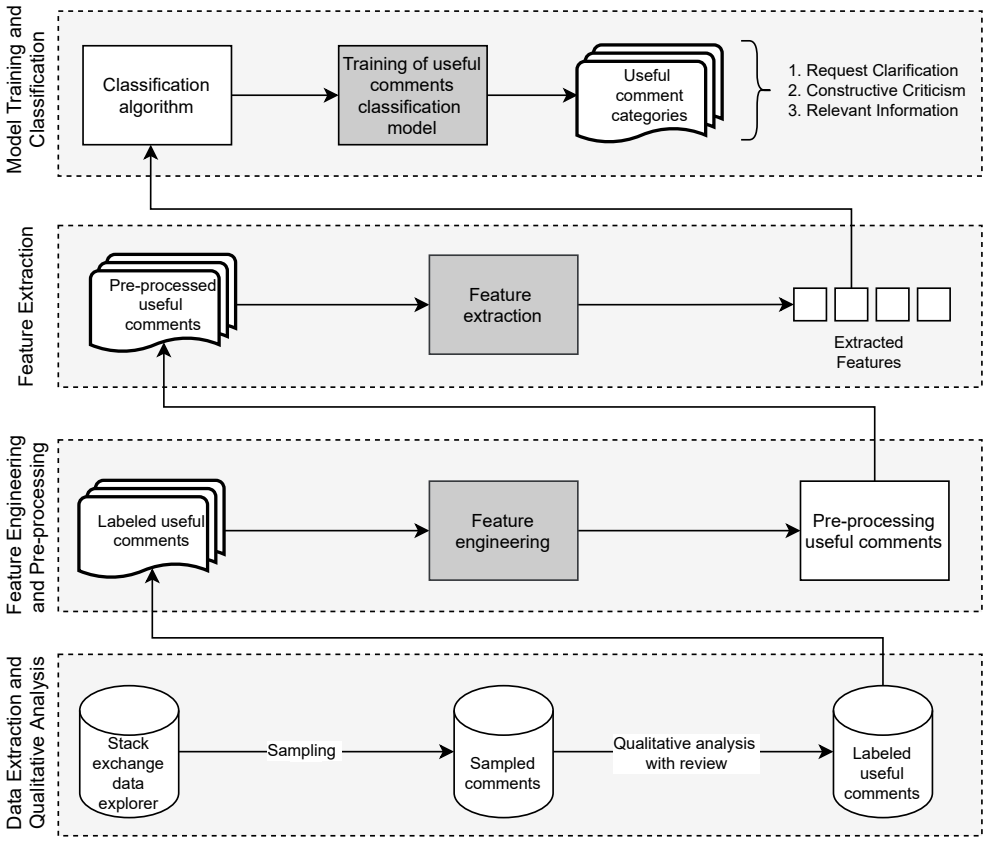


Fig. 1. High Level Architecture Diagram of Exploration of Useful Comments in Stack Overflow

```
INNER JOIN Posts ON Comments.PostId = Posts.Id
WHERE  Comments.Score > 0 AND Comments.CreationDate
   <'2020-11-08 00:00:00' AND Comments.CreationDate>
        '2015-01-01 00:00:00' AND Posts.Tags !=' '
ORDER BY Score Desc
```

The Stack Exchange Data Explorer was queried to fetch the comments which were made in response to questions and answers, that contain tags and with a score which is greater than 0. Comments with the comment score which is greater than 0 was considered as a measure of importance of comments. The query utilized in obtaining the data is presented in Listing 1.

As a result of querying, 50000 comments in Stack Overflow were obtained. From the obtained 50000 comments, 6164 comments were sampled using the simple random sampling technique which is a probability sampling technique that utilized randomization. This technique was used as there was no prior knowledge related to the comments data and therefore each element contained an equal chance of getting selected for the purpose of being a part of the sample[2].

B. Qualitative Analysis and Review

The objective of Qualitative Analysis was to create the dataset that is required to train the machine learning model. In Qualitative Analysis, the unclassified comments data was labeled to filter out the useful comments. A total of 6164 comments which were the output from data extraction were taken for the Qualitative Analysis. As the first step of Qualitative Analysis, deductive coding was used to label unclassified data [24]. In deductive coding a predefined set of label values were used to assign label values to unclassified comments. If a given comment was assigned multiple labels, such comments were removed from the dataset. As the output of deductive coding 3587 useful comments and 2577 not-useful comments were identified. Then as the next step of Qualitative Analysis, useful comments were filtered out and categorized into the three categories mentioned in Table I.

Table II discusses the measures used to categorize and label the extracted useful comments into standard comment categories during the qualitative analysis. After completing the qualitative analysis it was necessary to review the labeling process of the comments in order to measure the consistency, quality and accuracy of the labeled data. The review was performed by reviewing all the labels of the 3587 useful comments. The intra rater reliability percentage was calculated because both the data labeling and review was performed by a single annotator. Therefore, as a result of the review 3120 comments were properly labeled with the intra rater reliability of 86.98% with regard to correct labeling of comments in both occasions. Therefore 467 comments were disregarded in the study further as they were identified as misclassified in the review. 3120 useful comments were identified from the review process and those comments were used in the implementation

[2]https://towardsdatascience.com/sampling-techniques-a4e34111d808

TABLE I. QUALITATIVE ANALYSIS RELATED TO USEFUL COMMENTS IN STACK OVERFLOW FOR LABELING OF USEFUL COMMENTS

Useful Comment Category	Property Description
Request Clarification (Zhang et al., 2019a) (Requesting clarification from the author)	Requesting provision of more information or Expression of lack of understanding. These comments can be identified with the keywords such as 'please clarify', 'please elaborate', 'how', 'what' etc.
Constructive Criticism (Zhang et al., 2019a) (Guiding the author in improving the post)	Contains both positive and negative comments that are stated in a pleasant manner. Areas of improvement of posts are stated. Formatting and indentation issues are included more often.
Relevant Information (Zhang et al., 2019a) (Relevant but minor or transient information)	These comments may include a link to a related post, a link that redirects to other websites. Statements about question updates and answer updates were rarely found.

of the classification model. Out of the 3120 useful comments 1010 comments were of Constructive Criticism, 1047 comments were of Relevant Information and 1063 comments were of Request Clarification.

C. Feature Engineering and Pre-Processing

Features are independent individual variables that act as an input to a machine learning model. Machine Learning models use features for making predictions. Therefore, feature engineering was performed which included feature creation and evaluation of the created features through histogram plots. Feature creation was needed to construct new features from the 3120 useful comments identified after the review process. Moreover, feature engineering was important to identify whether external features other than the textual features can be used in building the classification model. Textual features include text scores. II presents the new features which were created and evaluated. Overlapping of data was clearly identified through horizontal scaling of histograms plotted for all comment categories separately and as a whole.

TABLE II. MEASURES USED FOR THE CATEGORIZATION

Feature	Description
Comment Length	Total number of characters in the useful comments
Comment Score	The number of upvotes a specific useful comment obtained in Stack Overflow
Punctuation Percentage	Percentage of the total number of punctuation used in a specific useful comment
Average Word Count	The mean word count of a specific useful comment
Capitalization Usage	The total number of capital letters used in a specific useful comment
Stop Words Count	The total number of stop words used in a specific useful comment
Positive Sentiment Score	The probability of the sentiment of a specific useful comment to be positive
Negative Sentiment Score	The probability of the sentiment of a specific useful comment to be negative
Neutral Sentiment Score	The probability of the sentiment of a specific useful comment to be neutral
Normalized Compound Score	The sum of negative sentiment score, positive sentiment score neutral sentiment score of a specific useful comment which is then normalized and ranges between -1 and +1

NLTK's VADER which is a parsimonious rule based model and was used in calculating the Sentiment Score of the comments data. VADER calculates the sentiment score of a text in terms of positive sentiment score, negative sentiment score, neutral sentiment score and normalized compound score. Fig. 2 shows the feature evaluation with horizontal scaling comment length and comment score of useful comments. Fig. 3 depicts the feature evaluation with horizontal scaling for punctuation percentage and average word count of useful comments. Fig. 4 depicts the feature evaluation with horizontal scaling for the count of capital letters and stop words of useful comments. Fig. 5 shows the feature evaluation for positive sentiment score and negative sentiment score of useful comments. Fig. 6 depicts the feature evaluation for horizontal scaling for count of capital letters of useful comments. Fig. 7 depicts the feature evaluation for horizontal scaling for count of stop words. Fig. 8 and Fig. 9 presents the feature evaluation for positive sentiment score and negative sentiment score respectively. Fig. 10 shows the feature evaluation for neutral sentiment score and Fig. 11 shows the feature evaluation for compound score of useful comments. Since majority overlapping areas were obtained in 2 or all 3 classes, the features created in the feature engineering process were disregarded and not utilized in building the classification model.

Preprocessing was done to remove the noisy data that might be present in the identified useful comments. During the preprocessing stage the identified useful comments data was preprocessed using Natural Language Processing Techniques. In preprocessing lowercasing the data, replacing URLs with a keyword , punctuation removal, stop words removal, tokenization of data, stemming and lemmatization, removal of numbers, removal of emojis and emoticons in comments, and handling of chat words were carried out. Comments that will be classified as Relevant Information contain URLs. So, it was necessary to capture the URLs and hence replacement of URLs with the keyword 'link' was done without the removal of URLs. As useful comments are also developer discussions, they contained emojis and emoticons in them and it was necessary to remove them in the preprocessing stage. It was also required to replace the chat words with their meaningful phrases. This was done by maintaining a list of chat words and their meaningful phrases as key value pairs in a text file. This chat words document included chat words derived from the glossary dictionary of Stack Exchange and some common chat words observed during the qualitative analysis.

D. Feature Extraction

Feature extraction was done by using the preprocessed comment data. Feature extraction was performed to extract the textual features from the useful comments. In this step comment text i.e. words of each comment were taken as the

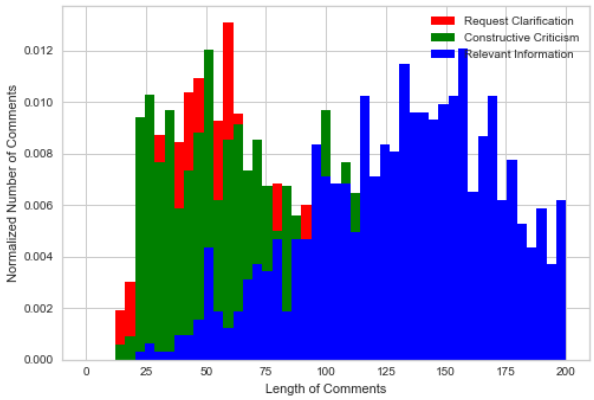


Fig. 2. Feature Evaluation with Horizontal Scaling for Useful Comment Length

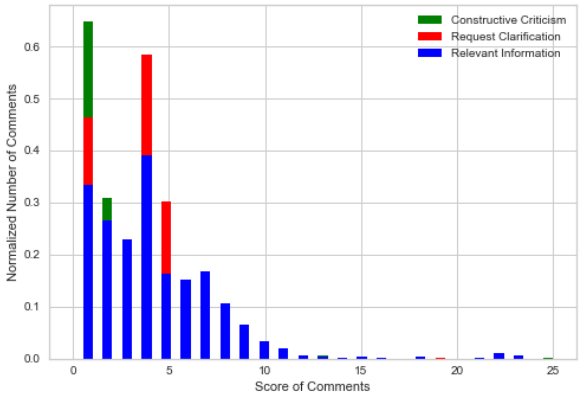


Fig. 3. Feature Evaluation with Horizontal Scaling for Useful Comment Score

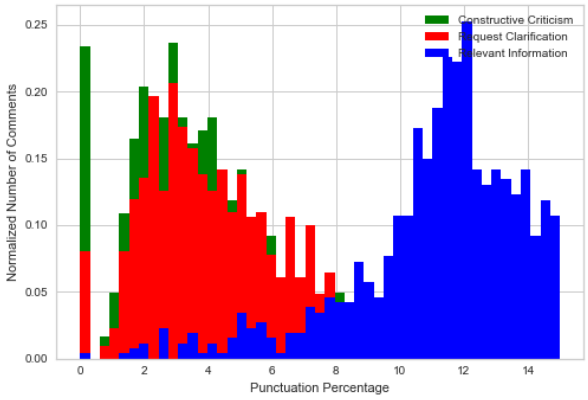


Fig. 4. Feature Evaluation with Horizontal Scaling for Punctuation Percentage of Useful Comments

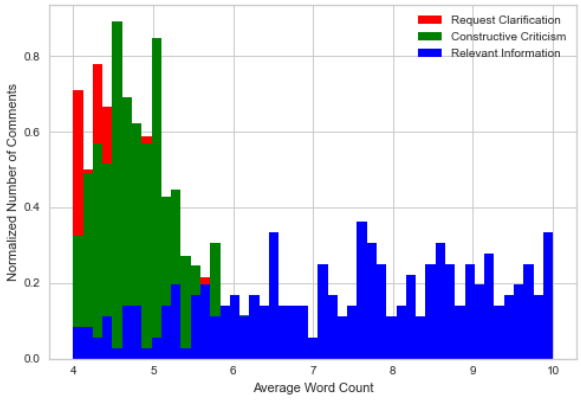


Fig. 5. Feature Evaluation with Horizontal Scaling for Average Word Count of Useful Comments

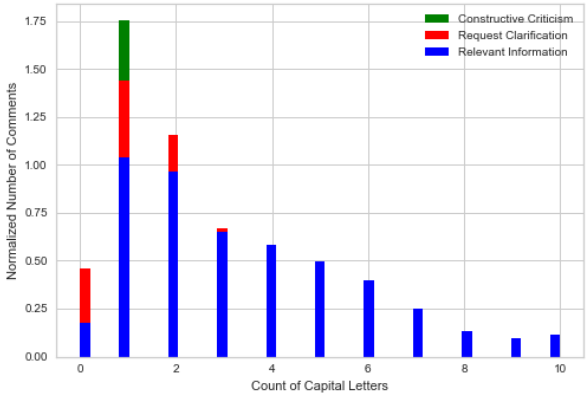


Fig. 6. Feature Evaluation with Horizontal Scaling for Count of Capital Letters of Useful Comments

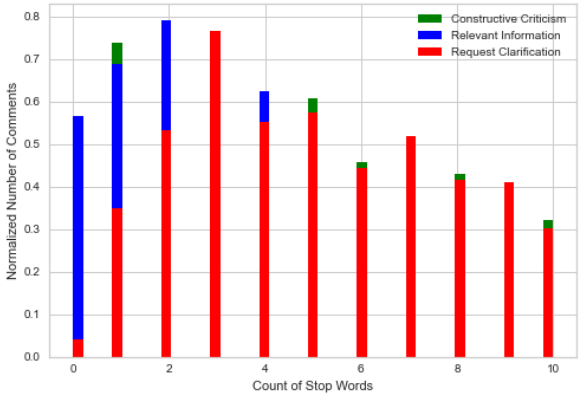


Fig. 7. Feature Evaluation with Horizontal Scaling for Count of Stop Words of Useful Comments

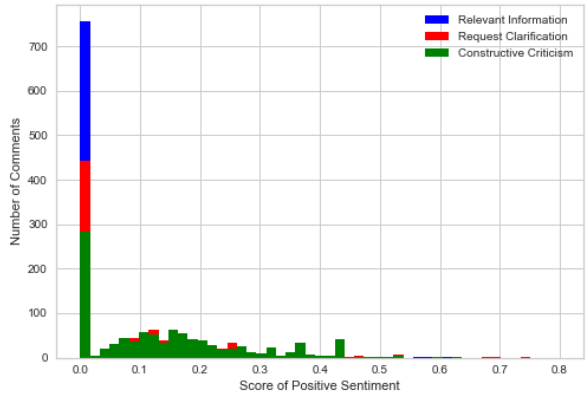


Fig. 8. Feature Evaluation for Positive Sentiment Score of Useful Comments

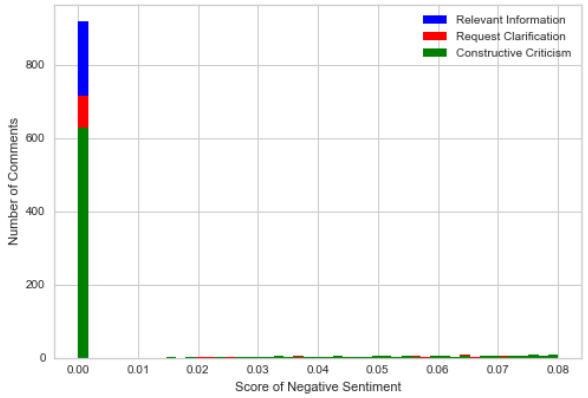


Fig. 9. Feature Evaluation for Negative Sentiment Score of Useful Comments

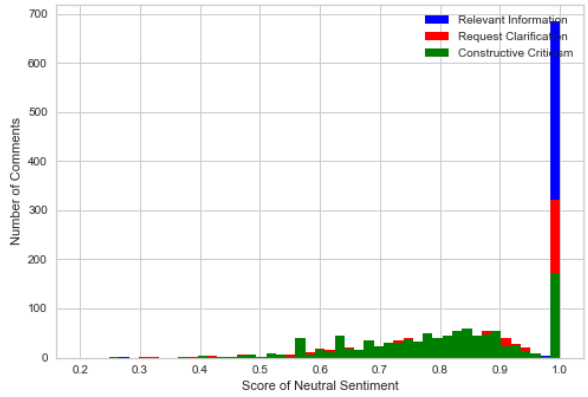


Fig. 10. Feature Evaluation for Neutral Sentiment Score of Useful Comments

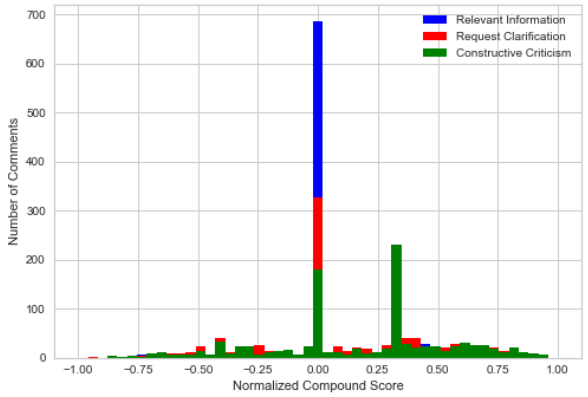


Fig. 11. Feature Evaluation for Compound Score of Useful Comments

features since the numeric features gained through feature engineering were exposed to overlapping of data in several comment categories. The TF-IDF feature extraction technique (Sulke & Varude, 2019) was utilized for the text feature extraction process along with N-grams. TF-IDF calculates the term frequency and inverse document frequency [25] [26]. IDF suppresses the effect of words which occurs in all 3 comment categories. Moreover, the TF-IDF vector looks the same as the word vector. This mechanism is used to find the meaning of sentences and it cancels out the incapableness of the bag of words feature extraction technique. TF answers how many times a particular word is used in the entire document. IDF calculates the importance of a certain term in a list of documents. For the feature extraction purpose TF-IDF Vectorizer was used as it performs the task of count vectorizer which is followed by TF-IDF transformer. Unigrams and Bigrams were utilized as N-grams. N-grams were used to boost the accuracy of the classification model. The N-gram frequency method provided an inexpensive and highly effective method of classifying documents. Encoding the categorical labels was done using the LabelEncoder before extracting features of useful comments. TF-IDF and N-grams scores were the extracted textual features. Shuffling of data was important to avoid the biases of data location within the data set. After the completion of feature extraction, sparse matrices with 275 features were obtained.

E. Training and Evaluating the Model

For the classification of useful comments, a multi-class SVM classifier was designed as the classification model. The features extracted in the previous step were fed into the designed classification model for the training purpose. 80% of data in the dataset was utilized for training the model and 20% was utilized for testing the model. The initial SVM classifier model was trained without any parameters. In this initial model a RBF kernel was used by default since the kernel was not specified.

Hyperparameter Tuning: Hyperparameters control the behaviour of the overall machine learning model. Therefore hyperparameter tuning was considered as a necessary step. The ultimate goal was to discover the optimal combination of hyperparameters of the SVM Model that minimizes the loss and maximizes the overall accuracy of the Model. Since certain

hyperparameter combinations are not supported with specific kernels in SVM, hyperparameters tuning was done separately for 5 SVM kernels which include Linear Kernel, RBF kernel, Polynomial kernel, Sigmoid kernel and Precomputed kernel. The objective was to identify the best Kernel with best hyperparameter combinations. For the Hyperparameter tuning sklearn's GridSearchCV was used. Since Cross Validation was done with GridSearchCV to obtain the best combination of hyperparameters, a validation dataset was not needed as the cross validation divided the training data set into k number of folds , and k-1 folds were used for the training purpose and the remaining fold was utilized as the validation set. In this hyperparameter tuning, cross validation with 3 folds was performed. C, kernel, degree, gamma, decision_function_shape, coefO were the utilized hyperparameters in hyperparameter tuning. Table III contains the optimal combination of hyperparameters obtained for each kernel. In SVM, gamma is the kernel coefficient, degree is the degree of the polynomial kernel function and coef0 is an independent term. When x,y are the data to be classified, the utilized kernels are as follows.

TABLE III. OPTIMAL COMBINATION OF HYPERPARAMETERS FOR EACH SVM KERNEL

SVM Kernel	Hyperparameters					
	C	kernel	degree	gamma	decision_ function_ shape	coef0
Linear	1	linear	-	1	ovo	-
RBF	2	rbf	-	scale	ovo	-
Polynomial	1	poly	1	scale	ovo	0.01
Sigmoid	10	sigmoid	-	0.1	ovo	0.01
Precomputed	1	precomputed	-	1	ovo	-

Linear Kernel: Mostly Used when a large number of data is available and when the data is linearly separable. It is the simplest Kernel Function in SVM. The Linear Kernel function (LK(x,y)) is denoted by the equation 1.

$$LK(x, y) = SUM(x.y) \tag{1}$$

RBF Kernel: It is usually used in classifying non-linear data. Proper separation of data when there is no prior knowledge about the data is performed successfully. The formula of RBF(RBFK(x,y)) is denoted by equation 2. Note that gamma varies between 0 and 1.

$$RBFK(x, y) = exp(-gamma||x - y||^2) \tag{2}$$

Polinomial Kernel: This kernel is a generalized representation of the Linear kernel. The formula of the polynomial kernel (PK(x,y)) is denoted by equation 3.

$$PK(x, y) = (gamma < x, y > +coefO)^{degree} \tag{3}$$

Sigmoid Kernel: This is often known as hyperbolic tangent or multilayer perceptron. This kernel is mostly preferred in Neural Networks. The formula of the sigmoid kernel (SK(x,y)) is as denoted by equation (4).

$$SK(x, y) = tanh(gamma < x, y > +coefO) \tag{4}$$

One-Vs-One(OVO) decomposition strategy was used in training the SVM model as it results in higher performance when compared with Non-OVO approaches, disregarding the overlapping level. OVO benefits the multi-class classification while increasing the separability of classes. Moreover, it was identified as an approach that highly benefits SVM as it provides robust results and superior performance [27]. Building and training of the SVM Models for linear, rbf, polynomial, sigmoid and precomputed kernels was done using the best combination of hyperparameters obtained through hyperparameter tuning. Identification of the best SVM Kernel that fits the problem context was done through evaluation measures such as the Holdout Method, Confusion Matrix and Classification Report.

IV. RESULTS AND EVALUATION

As the exploration of useful comments in Stack Overflow was based on building a classification model which categorizes useful comments with their respective useful comment Category, the evaluation of this model was carried out to evaluate classification accuracy, precision, recall, f1-score and the number of correctly classified instances of each class using the Holdout method, Confusion Matrix and the Classification Report. The test set contained 624 data entries. Before evaluation it was necessary at first to gain insight of the instances belonging to each category in test data. Therefore, data count of each comment category in the test set was obtained. Test set contained 202 instances of Request Clarification comments, 218 instances of Relevant Information comments and 204 instances of Constructive Criticism comments.

A. Evaluation of Classification Accuracy and F1-Score

In the Holdout Method the data set was divided into two parts, such as train set and test set. Train set contained 80% of the data while the Test set contained 20% of the data. The Train set was utilized to train the data and the Test set was utilized to test the predictive power of the implemented classification model. Classification Accuracy and F1-Score was gained as the metric of evaluation in the Holdout method. Holdout method based evaluation was performed to the initial SVM model which was built without any parameters or hyperparameters and also to the SVM models built and trained with distinct kernels and the optimal combinations of hyperparameters. Table IV contains the summary of the results obtained through the holdout method for the SVM models. According to the results obtained through the holdout method, the SVM model with the RBF kernel can be identified as the best classification model built and trained with optimal combination of hyperparameters as it promised the highest Accuracy of 87.02 and highest F1-Score of 87.11.

According to the results obtained through the holdout method, initial SVM Model and the SVM Model with RBF kernel promised similar accuracies which is 87.02. The initial SVM model promised the highest F1-score which is 87.21 when compared with all the other SVM models.

B. Evaluating the Number of Correctly Classified Instances of Each Class

For the purpose of gaining insights of the performance of the classification model the confusion matrix can be used. The

TABLE IV. SUMMARY OF THE RESULTS OF HOLDOUT METHOD FOR SVM MODELS BUILT WITH THE BEST COMBINATIONS OF HYPERPARAMETERS

Evaluation Metrics	Initial SVM Model	SVM Models with optimal combinations of hyperparameters				
		Linear SVM Model	RBF SVM Model	Polynomial SVM Model	Sigmoid SVM Model	Precomputed SVM Model
Accuracy	87.02	85.58	87.02	85.74	85.42	85.58
F1-Score	87.21	85.77	87.11	85.92	85.61	85.77

information gained from the confusion matrix can be used to determine the usefulness of the classification model. As a result important metrics such as accuracy, precision and recall can be determined. In this study, since useful comments fall into three categories the 3x3 Confusion Matrix was used for evaluation. In the 3x3 Confusion Matrix diagonal values were identified as correctly classified and non-diagonal values were identified as misclassified. The confusion matrix was drawn for the initial SVM Model which was built before hyperparameter tuning. Afterwards, the confusion matrices were plotted for each SVM Model with distinct kernels and optimal combination of hyperparameters obtained after hyperparameter tuning. The results obtained through the Confusion Matrices relevant to each SVM Model are summarized in Table V.

C. Evaluating the Quality of Predictions

The quality of the predictions relevant to a classification algorithm is measured by the Classification Report. It provides the main classification metrics based on each class. True Positives, True Negatives, False Positives and False Negatives are utilized for the purpose of predicting metrics in a Classification Report. Along with the Classification Report metrics relevant to macro average, micro average and weighted average were calculated for precision, recall and f1-score. Initially, the Classification Report was obtained for the initial SVM Model which was built without including any parameters. Afterwards, the Classification Reports were gained for each SVM Model with distinct kernels and optimal combination of hyperparameters. Table VI contains the summary of the results obtained from the Classification Report.

According to the evaluation mechanisms the initial SVM model which uses the RBF kernel in default has the highest weighted precision of 0.88 when compared to the rest of the SVM models. Among the SVM models trained with optimal combinations of hyperparameters, SVM model with the RBF kernel is the best kernel which classifies useful comments in stack overflow as it promised the highest accuracy, weighted average for recall and f1-score. Thus, research question is successfully addressed.

V. CONCLUSION

This paper presents a machine learning-based novel approach to explore Stack Overflow comments by classifying them into the respective standard comment categories. The main stages of this study consist of Data Extraction, Qualitative Analysis, Feature Engineering and Preprocessing, Feature Extraction, Training, and Evaluation of the classification model. As per the results of feature engineering, it was observed that none of the ten external features used can be combined with the textual features to implement the classification model. The evaluation was conducted to analyze the classification accuracy, precision, recall, f1-score, and the instances of each class that have been correctly classified by using the Holdout Method, Confusion Matrix, and Classification Report. RBF was identified as the best Kernel in exploring useful comments in Stack Overflow regardless of the use of hyperparameters while training the classification model. The results of this study can be utilized in the long process of Stack Overflow's useful comment analysis approaches for improving the facilitation of the process of learning and knowledge construction.

Future research may leverage the Multi-label classification approach to identify and classify the useful comments which belong to multiple comment categories.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the University of Moratuwa SRC Long-term Grant (Grant no: SRC/LT/2021/02).

REFERENCES

[1] S. Sengupta and C. Haythornthwaite, "Learning with comments: An analysis of comments and community on stack overflow," in *Hawaii: 53rd Hawaii International Conference on System Sciences*, 2020, pp. 2898 – 2907.

[2] A. Fontão, B. Ábia, I. Wiese, B. Estácio, M. Quinta, R. P. dos Santos, and A. C. Dias-Neto, "Supporting governance of mobile application developers from mining and analyzing technical questions in stack overflow," *Journal of Software Engineering Research and Development*, vol. 6, no. 1, pp. 1–34, 2018.

[3] A. Zagalsky, D. M. German, M.-A. Storey, C. G. Teshima, and G. Poo-Caamaño, "How the r community creates and curates knowledge: an extended study of stack overflow and mailing lists," *Empirical Software Engineering*, vol. 23, no. 2, pp. 953–986, 2018.

[4] M. M. Rahman, C. K. Roy, and I. Keivanloo, "Recommending insightful comments for source code using crowdsourced knowledge," in *The 15th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 2015)*, 2015, pp. 81 – 90.

[5] J. Cheriyan, B. T. R. Savarimuthu, and S. Cranefield, "Norm violation in online communities–a study of stack overflow comments," in *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*. Springer, 2017, pp. 20–34.

[6] A. Diyanati, B. S. Sheykhahmadloo, S. M. Fakhrahmad, M. H. Sadredini, and M. H. Diyanati, "A proposed approach to determining expertise level of stackoverflow programmers based on mining of user comments," *Journal of Computer Languages*, vol. 61, p. 101000, 2020.

[7] H. Zhang, S. Wang, T.-H. Chen, A. E. Hassan, and Y. Zou, "An empirical study of obsolete answers on stack overflow," in *The 42nd IEEE International Conference on Software Engineering*, 2020, pp. 1 – 14.

[8] P. Rani, S. Abukar, N. Stulova, A. Bergel, and O. Nierstrasz, "Do comments follow commenting conventions? a case study in java and python," in *2021 IEEE 21st International Working Conference on Source Code Analysis and Manipulation (SCAM)*. IEEE, 2021, pp. 165–169.

TABLE V. SUMMARY OF THE RESULTS OBTAINED THROUGH THE CONFUSION MATRICES RELEVANT TO EACH SVM MODEL

SVM Models		Number of correctly classified comments in each useful comment category			Total number of correctly classified comments
		Constructive Criticism	Relevant Information	Request Clarification	
Initial SVM Model		173	186	184	543
SVM Models with optimal combinations of hyperparameters	Linear Kernel	175	174	185	534
	RBF Kernel	179	185	179	543
	Polynomial Kernel	176	174	185	535
	Sigmoid Kernel	175	174	184	533
	Precomputed Kernel	175	174	185	534

TABLE VI. RESULTS OBTAINED THROUGH THE CLASSIFICATION REPORT

Metrics of Evaluation	SVM Models					
	Initial SVM Model	SVM Models with optimal combinations of hyperparameters				
		Linear Kernel	RBF Kernel	Polynomial Kernel	Sigmoid Kernel	Precomputed Kernel
Accuracy	0.87	0.86	0.87	0.86	0.85	0.86
Micro Average for Precision	0.87	0.86	0.87	0.86	0.85	0.86
Macro Average for Precision	0.88	0.87	0.87	0.87	0.87	0.87
Weighted Average for Precision	0.88	0.87	0.87	0.87	0.87	0.87
Micro Average for Recall	0.87	0.86	0.87	0.86	0.85	0.86
Macro Average for Recall	0.87	0.86	0.87	0.86	0.86	0.86
Weighted Average for Recall	0.87	0.86	0.87	0.86	0.85	0.86
Micro Average for F1-Score	0.87	0.86	0.87	0.86	0.85	0.86
Macro Average for F1-Score	0.87	0.86	0.87	0.86	0.86	0.86
Weighted Average for F1-Score	0.87	0.86	0.87	0.86	0.86	0.86

[9] A. Soni and S. Nadi, "Analyzing comment-induced updates on stack overflow," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 2021.

[10] S. Mondal, C. K. Saifullah, A. Bhattacharjee, M. M. Rahman, and C. K. Roy, "Early detection and guidelines to improve unanswered questions on stack overflow," in *14th Innovations in Software Engineering Conference (formerly known as India Software Engineering Conference)*, 2021, pp. 1–11.

[11] H. Tang and S. Nadi, "On using stack overflow comment-edit pairs to recommend code maintenance changes," *Empirical Software Engineering*, vol. 26, no. 4, pp. 1–35, 2021.

[12] A. Sulke and A. Varude, "Classification of online pernicious comments using machine learning," *International Journal for Scientific Research and Development, Oune, India*, vol. 7, no. 8, pp. 2321–0613, 2019.

[13] A. S. M. Venigalla, C. S. Lakkundi, and S. Chimalakonda, "Sotagger - towards classifying stack overflow posts through contextual tagging," 2019, pp. 493 – 496.

[14] S. Beyer, C. Macho, M. Di Penta, and M. Pinzger, "What kind of questions do developers ask on stack overflow? a comparison of automated approaches to classify posts into question categories," *Empirical Software Engineering*, vol. 25, no. 3, pp. 2258–2301, 2020.

[15] M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, "Classification of online toxic comments using the logistic regression and neural networks models," in *AIP Conference Proceedings*, 2018, pp. 1 – 5.

[16] M. Choetkiertikul, D. Avery, H. K. Dam, T. Tran, and A. Ghose, "Who will answer my question on stack overflow?" in *2015 24th Australasian Software Engineering Conference*, 2015, pp. 155 – 164.

[17] R. Abdalkareem, E. Shihab, and J. Rilling, "On code reuse from stackoverflow : An exploratory study on android apps," *Information and Software Technology*, vol. 88, 2017.

[18] Y. Wu, S. Wang, C.-P. Bezemer, and K. Inoue, "How do developers utilize source code from stack overflow?" *Empirical Software Engineering*, vol. 24, p. 637 – 673, 2019.

[19] G. Digkas, N. Nikolaidis, A. Ampatzoglou, and A. Chatzigeorgiou, "Reusing code from stackoverflow: The effect on technical debt," 2019.

[20] N. Novielli, F. Calefato, and F. Lanubile, "A gold standard for emotion annotation in stack overflow," 2018, pp. 14–17.

[21] E. Wong, J. Yang, and L. Tan, "Autocomment: Mining question and answer sites for automatic comment generation," in *IEEE, Palo Alto, USA*, 2013, pp. 562–567.

[22] C. Vassallo, S. Panichella, M. Di Penta, and G. Canfora, "Codes: Mining source code descriptions from developers discussions," *CODES: mining sourCe cOde Descriptions from developErs diScussions. Hyderabad*, 2014.

[23] N. Chakrabarty, *A Machine Learning Approach To Comment Toxicity Classification*, 2012.

[24] N. Pearse, "An illustration of deductive analysis in qualitative research," in *18th European Conference on Research Methodology for Business and Management Studies*, 2019, p. 264.

[25] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *International Journal of General Systems*, vol. 46, pp. 27–36, 2017.

[26] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, pp. 25–29, 2018.

[27] J. A. Sãez, M. Galar, and B. Krawczyk, "Addressing the overlapping data problem in classification using the one-vs-one decomposition strategy," pp. 83 396–83 411, 2019.

A New Index for Detecting Frequency of Unknown Underwater Weak Signals with Genetic Algorithm

Weixiang Yu, Xiukui Li
School of Information and Communication Engineering
Dalian University of Technology, China

***Abstract*—In this paper, a new index is proposed for detecting the frequency of unknown underwater signals based on the stochastic resonance theory. When the received weak signal is input into the stochastic resonance system, first, by frequency analysis, the frequency with the highest amplitude A_m of the output signal spectrum is considered as the pre-detection frequency. Then a cosine signal with the pre-detection frequency and unit amplitude is constructed. Define the pre-signal-to-noise-ratio as the logarithm of the squared amplitude A_m over the mean of signal amplitudes in all other frequencies. The new index is defined as the product of the pre-signal-to-noise-ratio and the correlation coefficient between the received unknown signal and the constructed cosine signal. The new index is featured by taking into account the signal characteristics in both time and frequency domain, and it will yield better signal frequency detection performance. In addition, to improve the time efficiency of the frequency detection, a method to bound the searching range, keyed to the genetic algorithm, of the stochastic resonance system parameters is proposed. The method can be used to detect the frequency of both single frequency and frequency-hopping unknown signals. With the designed new index and system parameter bounding method, the simulations and experiments for the weak underwater unknown signals are conducted. Compared to the piecewise mean value index and weighted power spectral kurtosis index, the new index yields a higher detection probability at varied input signal-to-noise ratios and signal frequencies. With bounding system parameter searching ranges, the time efficiency is improved. The main purpose of this paper is to detect the frequency of unknown underwater weak signals by stochastic resonance system with genetic algorithm. The main contributions are summarized as follows. First, the detection probability of weak signals is improved by stochastic resonance system with the proposed signal detection index than some other indexes. Second, to improve the time efficiency of the signal frequency detection, a method to bound the searching range of system parameters is proposed.**

***Keywords*—Stochastic resonance; underwater weak signal detection; genetic algorithm; frequency detection; frequency-hopping signal; index**

I. Introduction

The detection and identification of underwater unknown targets are of great significance for the coastal defense development. However, because of natural and human activities, the underwater environment is very complex, such as wind and waves on the sea surface, marine biological activities, ocean currents on the seabed and the movement of hulls. In addition, some underwater targets can change their frequencies and other information to hide themselves. These factors make the detection and identification of underwater targets more challenging. Therefore, the efficient detection of underwater targets in such a complex environment is very important for both scientific research and engineering practice. The methods of traditional weak signal detection usually use finite impulse response (FIR) and infinite impulse response (IIR) [1], [2] filters to filter out the background noise mixed with the signal. Although these methods have some effects on filtering the out-of-band noise of signal, they will fail when the noise is distributed in the signal band. Stochastic resonance (SR) [3] theory is a weak signal detection method with high efficiency, which is different from traditional signal detection methods. With the SR system, the weak signal can be enhanced, and the system output signal-to-noise ratio (SNR_o) will be maximized by utilizing the background noise. The SR system is mainly composed of weak signal, background noise, and nonlinear system [4]. SR theory was first proposed by R.Benzi *et al*. in 1981 [5], which explained a phenomenon that the glacial and warm climates occurred periodically in ancient climate. Then SR theory has been further developed. Nowadays, it has been applied extensively to many subjects such as meteorology [6], hydroacoustics [7], biomedicine [8] and mechanical mechanics [9], [10].

In recent years, researchers have paid more attention to the field of weak signal detection based on SR theory. Wang *et al*. The author in [11] extend the bistable SR system to the tri-stable system by adjusting the three potential heights of SR system potential function. They use the differential evolutionary particle swarm algorithm to search the optimal values of system parameters, which can effectively improve the system SNR_o by simulations and experiments. Zheng *et al*. The author in [12] propose a fractional-order stochastic resonance (FOSR) multi-parameter optimization algorithm based on genetic algorithm (GA), which is beneficial to the application and popularization of FOSR in weak signal detection. H. T. Reda *et al*. The author in [13] discuss the application of SR in spectrum sensing and propose a firefly-inspired algorithm to optimize the SR and noise parameters of the dynamic system to improve signal detection. Guo *et al*. [14], [15] detect the multi-frequency weak signal by the cascading and paralleling of SR system. Based on the adiabatic approximation theory [16], SR theory is applicable to the signal detection of low frequency ($\ll$1Hz). However, most of the signal frequencies in practical applications are not low. Leng *et al*. The authors in [17], [18] propose to transform high frequency to low frequency signals by using scale transformation and secondary sampling theory such that the high frequency weak signals can be detected based on SR theory. Ji *et al*. [19] realize the detection of hydroacoustic high frequency chirp signals by SR system, and verify the correctness and feasibility of this method by simulations and experiments. The performance of SR system is directly related to the detection probability

of weak signal. It is crucial to design a index with high adaptability for weak signal detection, which can keep the resonance response to follow the unknown signal features. The SNR_o can be increased by SR system, hence SNR_o and signal-to-noise ratio gain (SNR_g) [20], [21] are the most common and intuitive indexes for measuring the performance of the adaptive SR system. However, to calculate SNR_o and SNR_g, some prior information should be available, such as the frequency of weak signal. Normally, these prior information cannot be obtained in practical applications. It is indispensable to define a new index for the detection of completely unknown signals. Therefore, many new indexes have been proposed, such as weighted kurtosis index [22], time domain correlated kurtosis index [23], correlation coefficient index [24], weighted power spectrum kurtosis (WPSK) index [25], entropy kurtosis variation product index [26] and piecewise mean value index (PMV) [27]. Although these indexes are feasible in some conditions, their adaptability is constrained. For example, for some single indexes, the time and frequency domain characteristics of the signal cannot be taken together to detect the signal. For some multiple indexes, they are sensitive to signal frequency and pulse signal. When the input signal frequency is high or the input signal contains a large amount of pulse signals, with these indexes the weak signal detection probability is low. In this paper, a new index is proposed to improve the weak signal detection probability by SR system.

The contributions of this paper are as follows. First, a new index is designed for weak signal detection and GA is used as the optimization method to search the optimal values of system parameters. The signal detection probability with the new index is proved to be higher than that by SR systems with PMV and WPSK indexes through simulations and experiments. The designed new index takes into account both the time and frequency domain characteristics of the signal, and it is insensitive to the change of signal frequency. Second, to decrease the frequency detection time for both single frequency and frequency-hopping signals, a method to bound the searching ranges of system parameters is proposed. However, there are some limitations in the existing work. The experimental site is located at the seaside of Xing-hai Park in Dalian City in this paper, which is different from the deep sea. In addition, the simulation is carried out in Gaussian noise, which is different from the actual noise type.

The remaining parts of this paper are organized as follows. The SR theory and the definition of a new index are introduced in Section II. For single frequency and frequency-hopping signals, the method to bound the searching ranges of system parameters a and b is presented in Section III. The simulations and experiments are analyzed in Section IV and V. The results and discussion are introduced in Section VI. The future plans and improvements are introduced in Section VII. Finally, this paper is concluded in Section VIII.

II. New Index Definition and Design of SR System

A. Stochastic Resonance Theory

The bistable SR system can be given by nonlinear Langevin equation [28],

$$\frac{dx}{dt} = -V'(x) + s(t) + n(t) \tag{1}$$

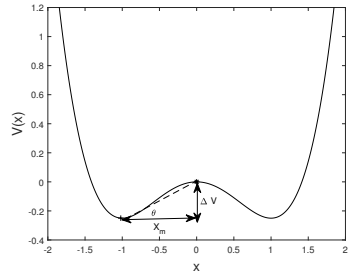


Fig. 1. The Potential Function Curve of $V(x)$, when a = b = 1. ΔV is the Potential Height, x_m is the Half Width of Potential well, $\tan\theta = \Delta V/x_m$.

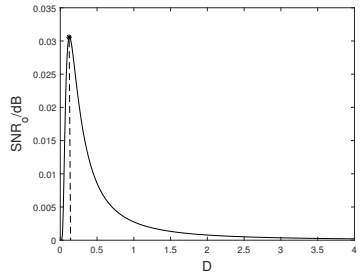


Fig. 2. The Curve of SNR_o with Noise Intensity D, when A = 0.1, a = b = 1.

where $V(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4$ is the potential function and its curve is shown in Fig. 1. x is the SR system output signal. System parameters a and b are real numbers greater than zero. $s(t) = A\cos(2\pi f_0 t)$ is the weak periodic signal to be detected, A is the signal amplitude and f_0 is the frequency. $n(t)$ is the noise, and $n(t) = \sqrt{2D}\varepsilon(t)$. D is the noise intensity and $\varepsilon(t)$ represents the Gaussian white noise with zero mean and variance one.

The SR system is in a stable state and the particle is in one of the two potential wells when there is no external signal input to the system. When an appropriate noise is input to the system, the particle will obtain energy from the noise and then skip the potential barrier to complete the transition between the two potential wells. Because the voltage difference between the two bistable potential wells is much larger than the original input signal amplitude, the SNR_o of SR system is greatly improved in this situation. The SNR_o of SR system is given by

$$SNR_o \approx \sqrt{2}\Delta V \left(\frac{A}{D}\right)^2 e^{-\frac{\Delta V}{D}} \tag{2}$$

where A is the amplitude of system input signal $s(t)$, ΔV is the potential height of the potential function. The SNR_o with respect to the noise intensity D is shown in Fig. 2.

It is inefficient and lacks of adaptivity to put the nonlinear system in a resonant state by adjusting noise intensity [29]. The potential height ΔV determined by the system parameters a and b is crucial for the particle to complete the periodic transition between the two potential wells. Parameters a and b can be searched by some optimization methods. In this paper, we choose GA as the optimization method to search

the optimal values of system parameters dynamically [30].

To detect the signal frequency by SR system, it is required that the input signal be of low frequency. A signal with high frequency can be down-converted into a low frequency signal by re-sampling the original signal with a ratio R. Then the low frequency $f_0^{'}$ of weak signal can be detected by SR system and finally we can restore the actual signal frequency by $f_0 = f_0' \times R$.

B. Definition of the New Index

For the frequency detection of unknown signals, various indexes have been proposed such as PMV and WPSK indexes. However, the signal frequency detection probability with these indexes are low. Herein, we propose a new index for the signal frequency detection through a SR system. Denote the received unknown weak signal by $r(t)$ and $r(k)$ is its sampled one signal. The new index is designed as follows.

I). The output signal $x(k)$ of SR system can be calculated by the fourth-order Longe-Kutta algorithm. The Fourier transform of $x(k)$ is denoted by $H(f)$. The frequency with the maximum amplitude of $H(f)$ is denoted by f_m.

II). Construct a new cosine signal $s^{'}(t) = \cos(2\pi f_m t)$ with frequency f_m and initial phase zero. The absolute value of the correlation coefficient between $r\,(t)$ and $s^{'}\,(t)$ is written as $C = |R(r(k), s^{'}(k))|$, where $r(k)$ and $s^{'}(k)$ are the sampling signals of $r\,(t)$ and $s^{'}\,(t)$, respectively.

III). The pre-signal-to-noise ratio of the output signal $x(k)$ is defined as $PSNR = 10 \times \log_{10} \frac{H^2(f_m)}{\sum_{f \neq f_m} H(f)/N}$, where $H\,(f_m)$ is the peak amplitude of $H(f)$ and N is the length of output signal $x(k)$.

IV). The new index, called CSNR, is defined as the product of C and $PSNR$, and the value of CSNR can be calculated by $\gamma_{csnr} = C \times PSNR$, which will be used as the fitness function of GA for searching the optimal values of system parameters a and b.

C. Process Detecting the Frequency of Weak Signals with the New Index

The sampled signal $r(k)$ of received unknown weak signal $r\,(t)$ is input to the optimal SR system and its spectrum $H(f)$ is obtained. The frequency with the maximum spectrum amplitude is considered as the frequency of $r\,(t)$. The frequency detecting process of the optimal SR system for detecting the frequency of weak signals is shown in Fig. 3.

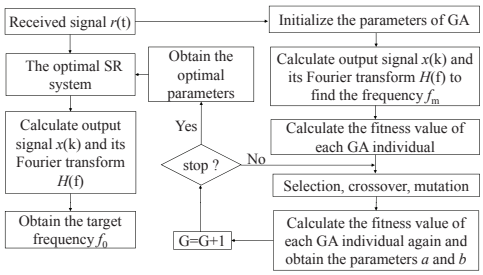


Fig. 3. The Frequency Detecting Process of SR System.

When the frequency with the maximum spectrum amplitude is not the target signal frequency f_0, $s^{'}\,(t)$ and $r\,(t)$ will not be highly correlated, and the value of C will be smaller. Hence, C can be used as a single time domain index. In addition, when the value of $PSNR$ is larger, the peak amplitude of output signal spectrum is larger than other amplitudes obviously, which can reduce the influence of other random frequencies on the frequency with the maximum spectrum amplitude. Hence, the value of $PSNR$ can be used as a single frequency domain index. To jointly consider both the time and frequency domain characteristics of the output signal, the product of the two single indexes can be a new index, which will yield better detection performance. When $n\,(t)$ is zero mean Gaussian noise, the correlation coefficient between $r\,(t)$ and $s^{'}\,(t)$ is

$$C = |R\,(r\,(k)\,, s'\,(k))| = \frac{|Cov\,(r\,(k)\,, s'\,(k))|}{\sqrt{Var\,(r\,(k)) \times Var\,(s'\,(k))}} \quad (3)$$

The denominator of (3) is a constant and its numerator is

$$\begin{aligned} c &= \left| \sum_{k=1}^{N} s(k) \times s^{'}(k) + \sum_{k=1}^{N} n(k) \times s^{'}(k) \right| \\ &= \begin{cases} c_1, f_m = f_0 \\ c_2, f_m \neq f_0 \end{cases} \end{aligned} \quad (4)$$

Generally, $c_1 \geq c_2$, where c is the numerator of (3), $s(k)$, $s^{'}(k)$ and $n(k)$ are the sampled signal of $s(t)$, $s^{'}(t)$ and $n(t)$, respectively. N is the number of signal $s(k)$. When $f_m = f_0$, $c = c_1$, otherwise, $c = c_2$.

III. A Method to Bound the Searching Ranges of System Parameters a and b

A. The Searching Range of System Parameter b for Single Frequency Signal Detection

The intensity of noise added to the nonlinear system will affect SR system operation. If the noise intensity is too low, the particle cannot obtain enough energy from the noise to skip the potential barrier. If the noise intensity is too high, although the particle can obtain enough energy from the noise to skip the potential barrier and resonance will occur, there will be a large amount of random noise mixed with the output signal. In this situation, the SNR_o of SR system is still very low and the signal frequency cannot be detected from the output signal with the strong background noise. Different potential heights ΔV determined by system parameters a and b have different fitness values of GA. The optimal values of system parameters a and b can be obtained by GA with the CSNR index. To reduce the searching time of system parameters a and b, we propose a method to limit the searching range of parameter b for GA.

From Fig. 1, the half width of potential well $x_m = \sqrt{\frac{a}{b}}$ and the potential height $\Delta V = \frac{a^2}{4b}$ [31]. The critical conditions for the particle to skip the potential barrier are given by

$$\begin{cases} \frac{\partial V(x,t)}{\partial x} = -ax + bx^3 = 0 \\ \frac{\partial^2 V(x,t)}{\partial x^2} = -a + 3bx^2 = 0 \end{cases} \quad (5)$$

Where x is the system output signal. Hence, the critical amplitude of particle is $A_c = \sqrt{\frac{4a^3}{27b}}$, which is considered as the threshold for the particle to skip the potential barrier. The amplitude of signal and noise intensity need to satisfy

$$\begin{cases} A \leq A_c \\ A_c \leq A + D \end{cases} \tag{6}$$

When $A>D$, the input signal-to-noise ratio (SNR_i) of system is high, and we can obtain the signal frequency from the received signal spectrum easily. Hence, assume $A \leq D$, from (6), we can obtain $b \geq \frac{4a^3}{27(A+D)^2} \geq \frac{4a^3}{27(2D)^2}$, where A is the amplitude of input signal $s(t)$, and D is the noise intensity. When the standard deviation of noise $n(t)$ is $\sqrt{2D}$, hence, we can obtain the received signal power $P \approx 2D$. Therefore, $b \geq \frac{4a^3}{27P^2}$ at low SNR_i, which indicates the potential height ΔV determined by the system parameters a and b would not be large. In addition, ΔV would not be small, otherwise, there will be a large amount of random noise mixed with the output signal. In this situation, the frequency f_0 of the target signal cannot be distinguished from other random frequencies. Hence, we can set a searching range for the parameter b. Define $\tan\theta = \frac{\Delta V}{x_m} = \frac{a\sqrt{a}}{4\sqrt{b}}$ (see Fig. 1 for θ). Therefore, we can obtain the searching range of parameter b, where P is the power of the received signal

$$\frac{4a^3}{27P^2} \leq b \leq \frac{a^3}{16\left(\tan\theta\right)^2} \tag{7}$$

B. The Searching Ranges of System Parameters a and b for Frequency-Hopping Signal Detection

To detect the frequency of frequency-hopping signal by the proposed SR system and reduce the searching time of parameters a and b, we will constrain the searching ranges of a and b. The Kramers rate r_k is the twice of signal frequency f_0 when resonance occurs [32].

$$r_k = \frac{a}{\sqrt{2}\pi r} e^{-\frac{a^2}{4bD}} = \frac{a}{\sqrt{2}\pi r} e^{-\frac{\Delta V}{D}} = 2f_0 \tag{8}$$

where r is the damping factor of the second order duffing equation in (9) and r_k is the Kramers rate.

$$\frac{d^2x}{dt^2} - r\frac{dx}{dt} = -V^{'}\left(x\right) + s\left(t\right) + n\left(t\right) \tag{9}$$

where x is the system output signal, $s(t)$ is the input signal, and $n(t)$ is the background noise.

The potential height ΔV affects the transition of the particle between the two potential wells and then the accuracy of signal frequency detection. Derived from (8), the relationship between the frequency f_0 and potential height ΔV is (10). Hence, ΔV needs to change while the input signal frequency changes. ΔV should decrease to help the particle to complete the transition between the two potential wells when the signal frequency f_0 increases. When the signal frequency f_0 decreases, ΔV should increase to reduce the speed of particle transition between the two potential wells, which can make the output signal frequency decrease and let it equal to the weak input signal frequency.

$$f_0 = \frac{a}{2\sqrt{2}\pi r} e^{-\frac{\Delta V}{D}} \tag{10}$$

The relationship between the values of two adjacent signal frequencies is $\frac{f_b}{m} < f_c < m \times f_b$, where m is a real number greater than zero. f_b is the signal frequency before the frequency changes, and the current frequency is f_c. Hence,

$$\frac{f_b}{f_c} \in \left(\frac{1}{m}, m\right) \tag{11}$$

Set $\Delta V_n = \frac{\Delta V_c - \Delta V_b}{D}$, where ΔV_b is the optimal potential height before the frequency changes and ΔV_c is the optimal potential height for the current frequency. Hence,

$$\Delta V_n \in \left(\ln\left(\frac{a_c}{m \times a_b}\right), \ln\left(\frac{m \times a_c}{a_b}\right)\right) \tag{12}$$

where a_b is the optimal value of parameter a before the frequency changes and a_c is the optimal a for the current frequency. When f_0 increases,

$$\begin{cases} \Delta V_n \in \left(\ln\left(\frac{a_c}{m \times a_b}\right), 0\right), \ \frac{a_b}{m} \leq a_c < m \times a_b \\ \Delta V_n \in \left(\ln\left(\frac{a_c}{m \times a_b}\right), \ln\left(\frac{m \times a_c}{a_b}\right)\right), \ a_c < \frac{a_b}{m} \\ 0, \ a_c \geq m \times a_b \end{cases} \tag{13}$$

When f_0 decreases,

$$\begin{cases} \Delta V_n \in \left(0, \ln\left(\frac{a_c}{m \times a_b}\right)\right), \ \frac{a_b}{m} \leq a_c < m \times a_b \\ \Delta V_n \in \left(\ln\left(\frac{a_c}{m \times a_b}\right), \ln\left(\frac{m \times a_c}{a_b}\right)\right), \ a_c \geq m \times a_b \\ 0, \ a_c < \frac{a_b}{m} \end{cases} \tag{14}$$

Generally, when $\frac{a_b}{m} \leq a_c < m \times a_b$

$$\begin{cases} \Delta V_n \in \left(\ln\left(\frac{a_c}{m \times a_b}\right), 0\right), \ \text{if} \ \ f_0 \ \text{increases} \\ \Delta V_n \in \left(0, \ln\left(\frac{a_c}{m \times a_b}\right)\right), \ \text{if} \ \ f_0 \ \text{decreases} \end{cases} \tag{15}$$

Hence, when f_0 increases,

$$\Delta V_c \in \left(\Delta V_b + D \times \ln\left(\frac{a_c}{m \times a_b}\right), \Delta V_b\right) \tag{16}$$

when f_0 decreases,

$$\Delta V_c \in \left(\Delta V_b, \Delta V_b + D \times \ln\left(\frac{m \times a_c}{a_b}\right)\right) \tag{17}$$

Because $\Delta V_c = \frac{(a_c)^2}{4b_c}$, where b_c is the optimal system parameter b for the current frequency. Hence,

$$\begin{cases} b_{c.min} = \frac{(a_c)^2}{4\Delta V_{c.max}} \\ b_{c.max} = \frac{(a_c)^2}{4\Delta V_{c.min}} \end{cases} \tag{18}$$

where $\Delta V_{c.min} = \Delta V_b + D \times \ln\left(\frac{a_c}{m \times a_b}\right)$, $\Delta V_{c.max} =$

$\Delta V_b + D \times \ln\left(\frac{m \times a_c}{a_b}\right)$. Therefore, $b_c \in (b_{c.min}, b_{c.max})$.

IV. Simulations

For the input signal, we set the amplitude $A = 0.1$ and frequency f_0 = 0.01 Hz. The signal sampling frequency f_s = 10 Hz, and the number of signal sample is $l = 10000$. The population size of genetic algorithm M = 100, crossover probability $p_s = 0.6$, and mutation probability $p_m = 0.001$. The searching ranges of system parameters a and b are $[0.0001 \sim 20]$ and $[0.0001 \sim 1000]$, respectively.

First, the background noise is assumed to follow zero mean Gaussian distribution with standard deviation $\sqrt{2D}$, and let $D = 1$ to verify that the SR system can improve the SNR_o. The time domain and frequency domain diagrams of the received signal and output signal are shown in Fig. 4. From Fig. 4. (a), when the noise intensity $D = 1$, it is observed that, due to the noise, the received signal appears non-periodic although the pure original signal is periodic. The periodicity of received signal $r(t)$ from the time domain waveform is not obvious, and the original cosine signal $s(t)$ is completely submerged in the background noise. The spectrum amplitude of target frequency $f_0 = 0.01$ Hz is not the biggest and smaller than that of some other frequencies in Fig. 4. (c). Therefore, the target frequency f_0 could not be detected from the received signal $r(t)$. However the time domain waveform of the output signal has obvious periodicity in Fig. 4. (b), and the spectrum amplitude of target frequency f_0 is significantly higher than that of other frequencies in Fig. 4. (d). The SNR_i = -27.6647 dB and the SNR_o = -8.6511 dB. The SNR_o increases by 19.0136 dB. Therefore, the SNR_o of weak signal with strong background noise can be increased and the target frequency can be detected by SR system.

A. The performance of SR System with the CSNR Index

We compare the performance of SR systems with the CSNR, PMV, and WPSK indexes for weak signal frequency detection. With 500 trials, the different frequency detection probability p of SR systems with different indexes are shown in Fig. 5, denoted by p_c, p_p and p_w, respectively. The detection probability p decreases with SNR_i decreases. p_c is higher than p_p and p_w when the SNR_i ranges from -32 to -19 dB. This indicates that the CSNR index is more robust. p_c and p_w are approximate to 1 and higher than p_p when the SNR_i ranges from -19 to -16 dB. From Fig. 5, the SNR_o of SR system with the CSNR index is improved by around 2.3 dB and 6.1 dB compared to the output by PMV and WPSK indexes, respectively when detection probability p is around 0.9.

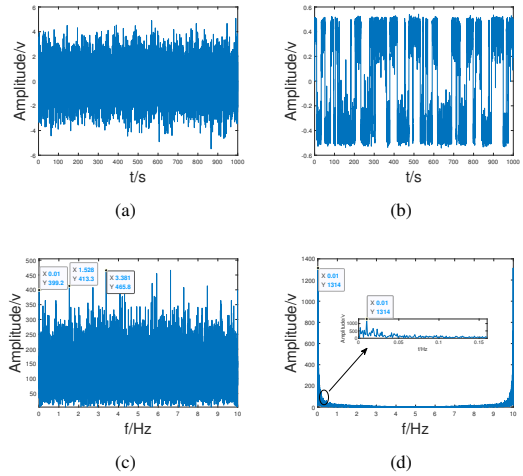


Fig. 4. (a) Received Signal in Time Domain, (b) Output Signal in Time Domain, (c) Received Signal in Frequency Domain, (d) Output Signal in Frequency Domain.

We compare the sensitivity of SR systems with the CSNR, PMV, and WPSK indexes, respectively, regarding the signal frequency change. Let noise intensity $D = 1$ and the input signal frequency f_0 varies from 0.01 to 0.3 Hz with step of 0.01 Hz. The detection results are shown in Fig. 6 and the detection probability p of SR systems with the three indexes decreases with the input signal frequency increases, which verifies that the SR system is adaptive to the detection of low frequency signal. p_c is higher than 0.8 and p_p is between 0.5 and 0.8. In addition, p_w is lower than 0.5 and decreases rapidly. Therefore, the SR system with CSNR index is less sensitive to the change of signal frequency and more robust than the SR systems with PMV and WPSK indexes.

B. Time Efficiency of Single Frequency Signal Detection

Let $\tan\theta = \Delta V/x_m = 1/30$ in Fig. 1. We compare the signal detection probability p and time T by SR systems with the different searching ranges of parameter b and different input signal-to-noise ratios. The detection results are shown in Fig. 7. It indicates that the detection probability p of weak signal frequency with different searching ranges of system parameter b decreases with the SNR_i. The detection probability p with different searching ranges of system parameter b are no obvious difference basically when SNR_i rangs from -32 to -16 dB. From Fig. 7. (b), it is clear that the detection time T with different searching ranges of system parameter b increases when SNR_i varies from -32 to -21 dB and then decreases when SNR_i varies from -21 to -16 dB. It indicates that when the signal amplitude $A = 0.1$, frequency $f_0 = 0.01$ Hz, and the

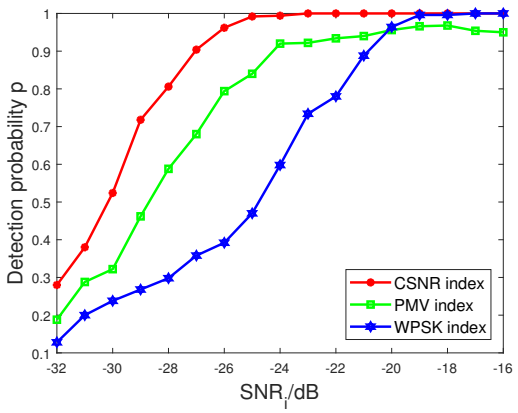


Fig. 5. The Single Frequency Signal Detection Probability by SR Systems with Different Indexes and Different Input Signal-to-Noise Ratios, where the Red, Green and Blue Curves are the Detection Probability with the CSNR, PMV and WPSK Indexes, respectively.

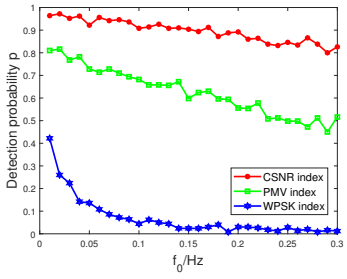
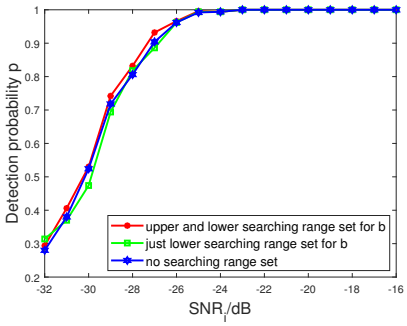
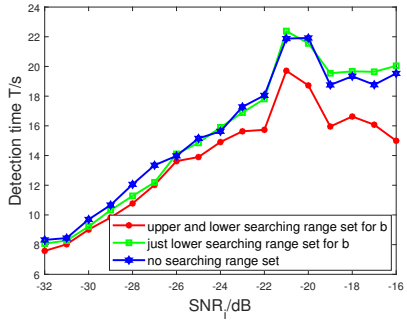


Fig. 6. The Single Frequency Signal Detection Probability by SR Systems with Different Indexes and Different Signal Frequencies, where the Red, Green and Blue Curves are Detection Probability with the CSNR, PMV, and WPSK Indexes, respectively.

SNR_i = -21 dB, the SNR_o of SR system will be maximized. The trends of three detection time curves are similar to the SNR_o with respect to the noise intensity D (see Fig. 2).



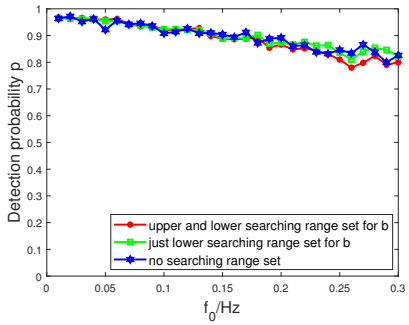
(a)



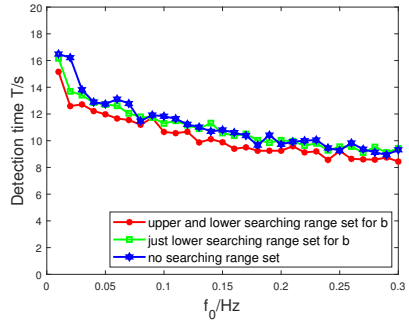
(b)

Fig. 7. With the Different Searching Ranges of Parameter b, the Single Frequency Signal Detection Probability and Detection Time with CSNR Index and Different Input Signal-to-Noise Ratios.

Denote the optimal noise intensity by D_o. When $D < D_o$, the SNR_o of SR system increases with noise intensity D. In this situation, when noise intensity D increases, the SR system needs more time to detect the weak signal. When $D = D_o$, the SR system has the maximal SNR_o and the detection time of signal frequency should be maximized. When $D > D_o$, the SNR_o of SR system decreases with the noise intensity D. The output signal of SR system is suboptimal and the detection time will decrease when noise intensity D increases. From Fig. 7. (b), the detection time with the upper and lower searching ranges of parameter b is less than that of the other searching ranges of parameter b obviously. Therefore, by bounding the searching range of parameter b, the detection time T can decrease. When the signal frequency changes with $D = 1$ and f_0 varies from 0.01 to 0.3 Hz with step of 0.01 Hz, the signal detection probability p and time T of SR systems with different searching ranges of parameter b are shown in Fig. 8. It is observed that the signal frequency detection time is decreased when the method of bounding the parameter b searching range is applied.



(a)

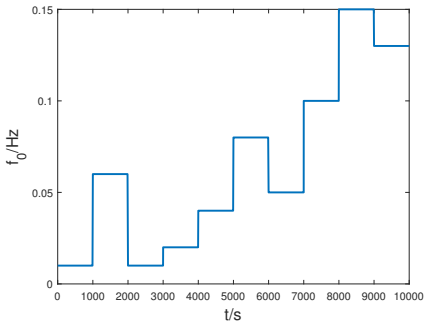


(b)

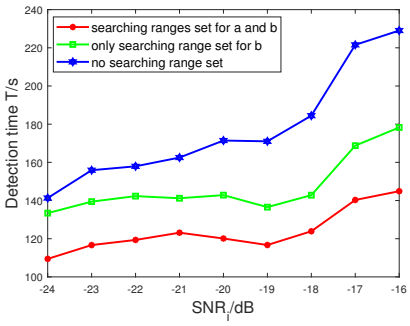
Fig. 8. With the Different Searching Ranges of Parameter b, the Single Frequency Signal Detection Probability and Detection Time with CSNR Index and Different Signal Frequencies.

C. Time Efficiency of Frequency-Hopping Signal Detection

For the input frequency-hopping signal, set amplitude A = 0.1, and the signal hopping frequency is [0.01, 0.06, 0.01, 0.02, 0.04, 0.08, 0.05, 0.10, 0.15, 0.13] Hz, respectively. The signal sampling frequency f_s = 10 Hz, and the number of sampling points of signal for each frequency is l = 10000. The change of signal frequency is shown in Fig. 9. (a). SNR_i ranges from -24 to -16 dB. With 100 trials, the signal detection time T with different searching ranges of parameters a and b is shown in Fig. 9. (b), and the detection probability p for each frequency is 1. It can be concluded that the detection time for frequency-hopping signal can decrease by SR system when the method of bounding the parameters a and b searching ranges is applied.



(a)



(b)

Fig. 9. (a) The Frequency of Original Frequency-Hopping Signal. (b) With the different Searching Ranges of Parameters a and b, the Frequency-Hopping Signal Detection Time by SR System with the CSNR Index.

V. Experiments

A. Single Frequency Signal Detection

The experimental site is located at the seaside of Xinghai Park in Dalian City. The signal of a motor with a fixed frequency is used as the weak signal to be detected. A mobile phone is used as the receiving device. The signal sampling frequency $f_s = 8000$ Hz. The motor and phone are placed below the sea surface around 0.5 m and the distance between them is around 2 m. The received signal r_s is shown in Fig. 10. r_s is amplified by a factor g, g = 3. We select three periods of signals r_i $(i = 1, 2, 3)$ from r_s. The number of signal samples for each period is $l = 10000$. The sample points are $[325000 \sim 335000]$, $[356000 \sim 366000]$, $[450000 \sim 460000]$ for r_1, r_2 and r_3, respectively. The detection results are shown in Tab. I and Fig. 11. The detected frequency $f_0^{'} = 0.028$ Hz by SR system and the actual signal frequency can be obtained by $f_0 = f_0^{'} \times R = 56$ Hz, where R is the ratio of down-converting the high frequency of actual signal into a low frequency, R=2000. From Fig. 11, the frequency of weak signal r_1 can be detected by SR systems with CSNR, PMV, and WPSK indexes, respectively. For the output signal, the spectrum amplitude of frequency $f_0^{'}$ with the CSNR index is higher than that by each of the other two indexes. From Tab. I, the SNR_o by SR system with the CSNR index is higher than the output by the other two indexes, and the detection time of single frequency by SR system with the upper and lower searching ranges of parameter b is less than that when no searching range is set for parameter b or only lower searching range is set.

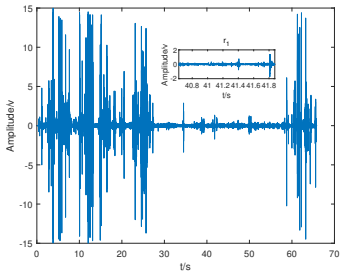


Fig. 10. Original Received Signal r_s

B. Frequency-Hopping Signal Detection

The signal of a motor with dynamic frequency change is used as the weak signal to be detected. The signal sampling frequency $f_s = 44100$Hz. The frequency of the motor is adjusted dynamically by a knob. The motor and phone are place below the sea surface around 0.5 m and the distance between them is around 4 m. The received signal r_s is shown in Fig. 12. r_s is amplified by a factor g, g = 200. We select ten periods of signals r_i $(i = 1...10)$ from signal r_s. The number of signal samples for each period is l = 10000. Then the ten periods of signals are combined into a new frequency-hopping signal s. From Fig. 13. (a) and Fig. 14. (a), the periodicity of signal s is not obvious. However, the output signals by SR systems with different parameters searching ranges show a strong periodicity (see Fig. 13. (b), (c), (d)). The time-frequency spectrum of the output signals by the SR systems with different searching ranges of system parameters a and b are shown in Fig. 14. (b), (c), (d), where the color brightness for each frequency in the signal spectrum indicates the magnitude of the normalized spectrum amplitude. In each of four cases, the target frequency is detected, i.e., the spectrum amplitude of the target frequency is the maximum and it is indicted by the brightest color. The frequency-hopping signal s is detected in four cases: 1) frequency-hopping signal s; 2) no searching ranges set for a and b; 3) set searching range of parameter b for single frequency signal; and, 4) set searching ranges of a and b for frequency-hopping signal. The detection results are the same for the four cases, however the periodicity of system output signal for the case 2), 3) and 4) are more strong than case 1). The detection frequencies are $f^{'}$ = [0.063945, 0.134505, 0.055125, 0.063945, 0.072765, 0.059535, 0.14112, 0.06615, 0.090405, 0.04851] Hz by SR systems with different searching ranges of parameters a and b, respectively. The actual signal frequencies can be obtained by $f = f^{'} \times R$ = [127.89, 269.01, 110.25, 127.89, 145.53, 119.07, 282.24, 132.30, 180.81, 97.02] Hz, where R=2000. The detection time of the frequency-hopping signal s with different searching ranges of parameters are 307.374s, 282.517s, and 226.736s, respectively. It indicates for frequency-hopping signal that the detection time by SR system by setting the searching ranges for a and b decreases by 80.638s and 55.781s, respectively, than that produced when no searching range is set or only the searching range for parameter b is set.

TABLE I. THE DETECTION RESULTS OF SINGLE FREQUENCY SIGNAL

Different indexes	r_1		r_2		r_3	
	SNR_o/dB	T/s	SNR_o/dB	T/s	SNR_o/dB	T/s
CSNR_{no}	-9.3987	16.7440	-8.2307	9.1730	-9.6321	9.5680
CSNR_{l}	-9.7140	10.6470	-8.5466	13.6340	-9.8430	18.5040
CSNR_{ul}	-9.3512	8.2720	-8.1565	8.5050	-9.6743	8.4370
PMV	-9.7326	15.1588	-8.6778	10.3360	-9.7843	8.7956
WPSK	-9.8321	13.7259	-8.8342	9.0453	10.3420	9.5723

[a]CSNR_{no} is the condition with the CSNR index and the no searching range set.
[b]CSNR_{l} is the condition with the CSNR index and the only lower searching range set for b.
[c]CSNR_{ul} is condition with the CSNR index and the upper and lower searching ranges set for b.

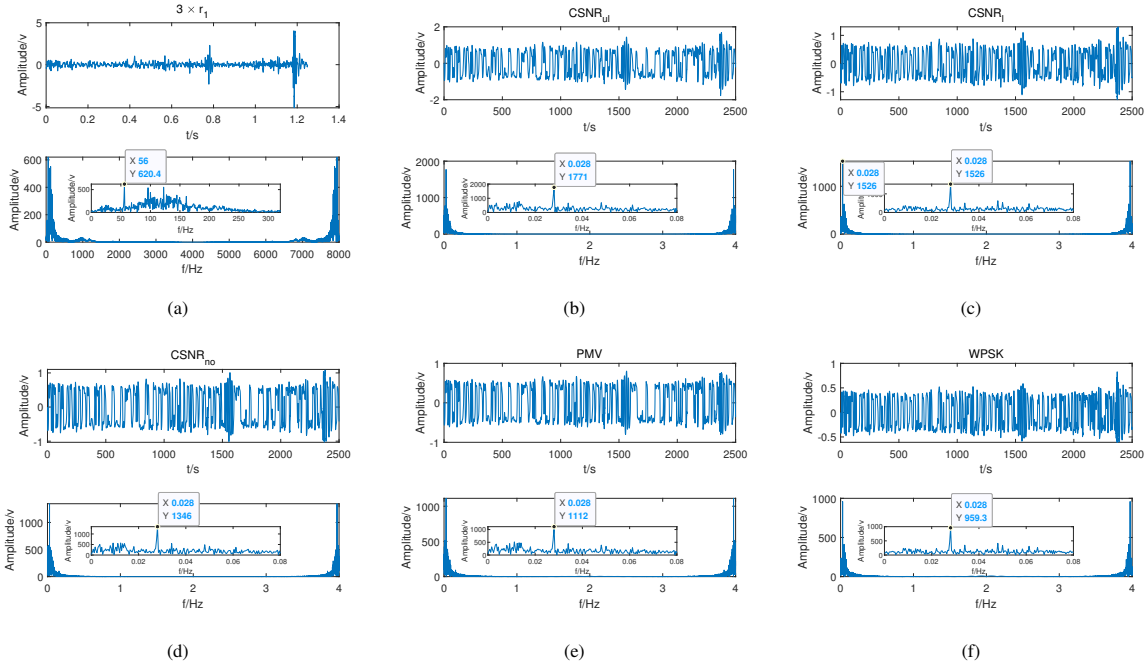


Fig. 11. The Detection Results of Single Frequency Signal r_1 by SR Systems with Different Indexes and Searching Ranges of Parameter b.

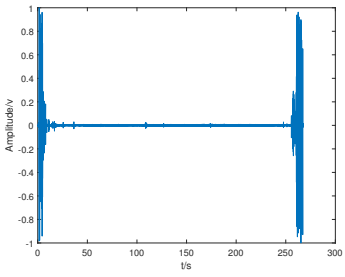


Fig. 12. Original Received Signal r_s.

VI. RESULTS AND DISCUSSION

The performance of weak signal detection by SR system with the CSNR index is verified by simulations and experiments. The frequency detection probability of SR systems with CSNR, PMV, and WPSK indexes are denoted by p_c, p_p and p_w, respectively. In simulations, p_c of weak signal with the CSNR index is 1 when the SNR_i is higher than -26 dB, and p_c is as high as 0.9 when SNR_i = -27.5 dB. p_c is higher than p_p and p_w when SNR_i varies from -32 to -19 dB. p_c and p_p are approximate to 1 and higher than p_w when SNR_i is $[-19 \sim -16]$ dB. When SNR_i = -26 dB, the performance of SR system with the CSNR index is insensitive to the change of input signal frequency. p_c is higher than 0.8 and p_p is between 0.5 and 0.8 when the input signal frequency f_0 varies from 0.01 to 0.3 Hz with step of 0.01 Hz. However, p_w is lower than 0.5 and p_w decreases rapidly. With the searching range set for system parameter b, the detection time of single frequency signal with the CSNR index decreases. The SNR_o of SR system with CSNR index is higher than the output by PMV and WPSK indexes in the same condition. For frequency-hopping signal, with the searching ranges set for system parameters a and b, when SNR_i varies from -24 to -16 dB, the signal detection time will decrease significantly. The experimental results are consistent with the simulation results.

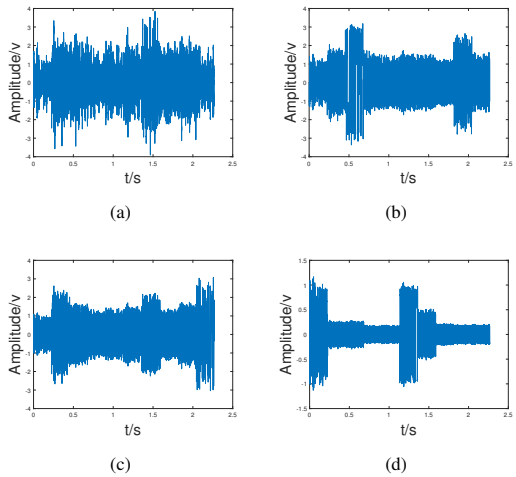


Fig. 13. The Signal in Time Domain. (a) Frequency-Hopping Signal s, (b) the Output Signal with no Searching Range Set, (c) the Output Signal with Searching Range Set for Parameter b only, (d) the Output Signal with Searching Ranges Set for Parameters a and b.

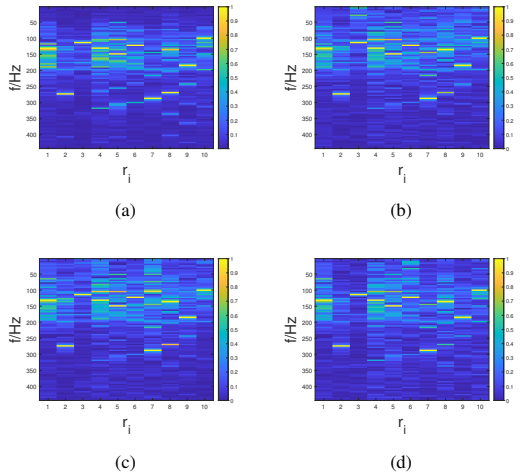


Fig. 14. The Signal in Time-Frequency Spectrum. The X-Label Represents Signal r_i $(i = 1...10)$ and Y-Label Represents the Output Signal Frequency (from 0 to 500 Hz). (a) Frequency-Hopping Signal s, (b) the Output Signal with no Searching Range Set, (c) the Output Signal with Searching Range Set for Parameter b only, (d) the Output Signal with Searching Ranges set for Parameters a and b.

VII. Future Plans and Improvements

The future plans are summarized as follows. First, the machine learning method can be combined with the genetic algorithm in this paper to further improve the detection time efficiency of unknown signals. Second, we can use underwater sonar array to improve the signal detection probability. The experimental verification part needs to be improved. I hope to conduct experiments in deep sea in the future to obtain more diverse and accurate underwater signals.

VIII. Conclusion

In this paper, a new index, called CSNR, is proposed to detect the frequency of unknown underwater signals based on SR theory with genetic algorithm, and the method of bounding searching ranges of system parameters a and b is presented to reduce the detection time. The performance of weak signal detection by SR system with the CSNR index is verified by simulations and experiments. The frequency detection probability of SR systems with CSNR, PMV, and WPSK indexes are denoted by p_c, p_p and p_w, respectively. The results show that p_c is higher than p_p and p_w. In addition, the performance of SR system with the CSNR index is insensitive to the change of input signal frequency. With the searching ranges set for system parameters a and b, the signal detection time will decrease significantly in simulations and experiments. In conclusion, the SR system with the proposed CSNR index and parameter searching ranges outperforms SR systems with PMV and WPSK indexes in terms of detection probability and detection time. Therefore, the contributions proposed in this paper are of positive significance to the detection of underwater weak signals in practical applications.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial or nonprofit sectors. My sincere acknowledgment goes to Dr. Li Xiukui.

References

[1] Li, X., Tan., W. Design and Implementation of IIR Multi-path Filter for SSVEP Based on MATLAB. 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), pp. 83–87, 2019.

[2] M, L., A., K., High speed FIR adaptive filter for RADAR applications. 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, pp. 2118–2122, 2016.

[3] Dong, H.,Wang, H., Shen, X., Huang, Z., Ma., S. Detection of underwater weak signal via matched stochastic resonance. OCEANS 2017 - Aberdeen, pp. 1–7, 2017.

[4] R. Lang, F.G. X. Li, Yang., L. Re-scaling and adaptive stochastic resonance as a tool for weak gnss signal acquisition. Journal of Systems Engineering and Electronics. 27(2), 290–296, 2016.

[5] Lucarini, V. Stochastic resonance for nonequilibrium systems. Phys. Rev.E. 100, 062124, 2019. https://doi.org/10.1103/PhysRevE.100.062124

[6] Bordi, I., Sutera., A. Stochastic perturbation in meteorology. Waves in Random Media. 10(3), 1–30, 2000. https://doi.org/10.1088/0959-7174/10/3/201

[7] Liu, W., Wang, Y., Liu, X. Weak thruster fault detection for auv based on stochastic resonance and wavelet reconstruction. Journal of Central South University. 23(11), 2883–2895, 2016.

[8] Han, D., li, P., An, S., Shi., P. Multi-frequency weak signal detection based on wavelet transform and parameter compensation band-pass multi-stable stochastic resonance. Mechanical Systems and Signal Processing. 70-71, 995–1010, 2015. https://doi.org/10.1016/j.ymssp.2015.09.003

[9] Xuefang Xu, Y.L. An incorrect data detection method for big data cleaning of machinery condition monitoring. Industrial Electronics. 67(3), 2326–2336, 2020.

[10] Shi, P., Yuan, D., Han, D., Zhang, Y., Fu., R. Stochastic resonance in a time-delayed feedback tristable system and its application in fault diagnosis. Journal of Sound and Vibration. 424, 1–14, 2018. https://doi.org/10.1016/j.jsv.2018.03.007

[11] Wang, Y., Jiao, S., Zhang, Q., Lei, S., Qiao., X. A weak signal detection method based on adaptive parameter-induced tri-stable stochastic resonance. CHINESE JOURNAL OF PHYSICS. 56, 1187–1198, 2018. https://doi.org/10.1016/j.cjph.2018.04.002

[12] Qi, W., Liu, Y., Guo, S., Wang, X., Guo., Z. An adaptive data detection algorithm based on intermittent chaos with strong noise background. Neural Comput and Applic. 32, 16755–16762, 2018. https://doi.org/10.1007/s00521-018-3839-9

[13] Reda, H.T., Mahmood, A., Diro, A. Firefly-inspired stochastic resonance for spectrum sensing in cr-based iot communications. Neural Comput and Applic. 32, 16011–16023, 2019. https://doi.org/10.1007/s00521-019-04584-0

[14] Liu, W., Liu, Z. Magnetic anomaly signal detection using parallel monostable stochastic resonance system. IEEE Access. 8, 162230–162237, 2020. https://doi.org/10.1109/ACCESS.2020.3020881

[15] Guo, W., Zhou, Z., Chen, C. Cascaded and parallel stochastic resonance for weak signal detection and its simulation study. 2016 Prognostics and System Health Management Conference (PHM-Chengdu), 2016, pp. 1-6, doi: 10.1109/PHM.2016.7819839, 2016.

[16] Liu, H.G., Liu, X.L., Yang, J.H. Detecting the weak high-frequency character signal by vibrational resonance in the duffing oscillator. Nonlinear Dyn. 89(4), 2621–2628, 2017. https://doi.org/10.1007/s11071-017-3610-2

[17] Chen, L., Zhang, Y., Feng, A., Xu, Z., Li, B., Shen., H. A new model of stochastic resonance used in weak signal detection. AppliedMechanics and Materials. 43, 229–232, 2010. https://doi.org/10.4028/www.scientific.net/AMM.43.229

[18] Liu, J., Leng, Y., Lai, Z., Fan., S. Multi-frequency signal detection based on frequency exchange and re-scaling stochastic resonance and its application to weak fault diagnosis. Sensors. 18(5), 1325–1325, 2018. https://doi.org/10.3390/s18051325

[19] Shu-Yao, J., Fei, Y., Ke-Yu, C., En, C. Application of stochastic resonance technology in underwater acoustic weak signal detection. OCEANS 2016 - Shanghai, 2016, pp. 1-5, doi: 10.1109/OCEANSAP.2016.7485567 2016.

[20] Gauthier, P.-A. et al., A. Acoustical inverse problems regularization: Direct definition of filter factors using signal-to-noise ratio. Journal of Sound and Vibration. 333(3), 761–773, 2014.

[21] Cheng, W., Xua, X., Ding, Y., Sun., K. Stochastic resonance in a single-well potential and its application in rolling bearing fault diagnosis. Review of Scientific Instruments. 91(6), 761–773, 2020.

[22] Li, J., Chen, X., He., Z. Adaptive stochastic resonance method for impact signal detection based on sliding window. Mechanical Systems and Signal Processing. 36(2), 240–255, 2013. https://doi.org/10.1016/j.ymssp.2012.12.004

[23] Xiang, J., Zhong, Y., Gao., H. Rolling element bearing fault detection using ppca and spectral kurtosis. Measurement. 75, 180–191, 2015. https://doi.org/10.1016/j.measurement.2015.07.045

[24] Morales., C.A. Comments on the mac and the nco, and a linear modal correlation coefficient. Journal of Sound and Vibration. 282(1), 529–537, 2004. https://doi.org/10.1016/j.jsv.2004.04.011

[25] Wang, J., He, Q., Kong, F. Adaptive multiscale noise tuning stochastic resonance for health diagnosis of rolling element bearings. Instrumentation and Measurement. 64(2), 564–577, 2015.

[26] Huang, Z., Wang, H., Dong, H., Jian, S. Detection of Underwater Weak Signals Based on Stochastic Resonance Multi-Measures Fusion.2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO), pp. 1-6, doi:10.1109/OCEANSKOBE.2018.8559075, 2018.

[27] Huang, D., Yang., J. Recovering an unknown signal completely submerged in strong noise by a new stochastic resonance method. Communications in Nonlinear Science and Numerical Simulation. 66, 156–166, 2018. https://doi.org/10.1016/j.cnsns.2018.06.011

[28] Wang, X., Huang. Adaptive stochastic resonance method based on quantum genetic algorithm and its application in dynamic characteristic identification of bridge gnss monitoring data. IEEE Access. 8, 113994–114009, 2020. https://doi.org/10.1109/ACCESS.2020.3002889

[29] Li, J., Li, M., Zhang., J. Rolling bearing fault diagnosis based on time-delayed feedback monostable stochastic resonance and adaptive minimumentropy deconvolution. Journal of Sound and Vibration. 401, 139–151, 2017. https://doi.org/10.1016/j.jsv.2017.04.036

[30] Yan Ren, J.H. Research on fault feature extraction of hydropower units based on adaptive stochastic resonance and fourier decomposition method. Shock and Vibration. 2021, 2021. https://doi.org/10.1155/2021/6640040

[31] Qiu, Y., Yuan, F., Ji, S., Cheng., E. Stochastic resonance with reinforcement learning for underwater acoustic communication signal. Applied Acoustics. 173, 2021. https://doi.org/10.1016/j.apacoust.2020.107688

[32] Li, J., Li, M., Zhang., J. Research on detection method of multi-frequency weak signal based on stochastic resonance and chaos characteristics of duffing system. CHINESE JOURNAL OF PHYSICS. 64, 333–347, 2020. https://doi.org/10.1016/j.cjph.2019.12.001

Extraction of Point-of-Interest in Multispectral Images for Face Recognition

Kossi Kuma KATAKPE
Institut de Mathématiques et de Sciences Physiques (IMSP)
Université d'Abomey Calavi (UAC)
Porto Novo, Benin

Lyes AKSAS
UFR Science & Technology
IEM Department, ImVia Laboratory, Color, Sensor & Multispectral Imaging
Burgundy University
Dijon, France

Diarra MAMADOU
UFR Mathématiques et Informatique Université Félix Houphouët-Boigny,
Abidjan, Côte d'Ivoire

Pierre GOUTON
UFR Science & Technology
IEM Department, ImVia Laboratory, Color, Sensor & Multispectral Imaging
Burgundy University
Dijon, France

***Abstract*—Security systems in companies, airports, enterprises, etc. face numerous challenges. Among the major ones there is objects or face recognition. The problem with the robustness of recognition systems that usually affects color images nowadays can be addressed by multispectral image acquisition in the near infrared range with cameras equipped with new high performance sensors able to take images in dark or uncontrolled environments with much more accuracy. Multispectral CMOS (Complementary Metal Oxide Semi-conductor) sensors in a single shot record several wavelengths that are isolated and allow very specific analyses. They are equipped with new acquisition methods and provide observations that are more accurate. The current generation of these imaging sensors involve scientific and technical interest because they provide much more information than those that operate in visible range; precise nature and spatio-temporal evolution of the areas need to be analyzed. In this study, multispectral images acquired by camera equipped with a hybrid sensor operating in near infrared has been used. This camera is built in the ImViA laboratory of the University of Bourgogne as part of the European project EXIST (EXtended Image Sensing Technologies). The process involved in image acquisition, image mosaicing and image demosaicing by using mosaic filters. After acquisition process the interest points be extract in these bands of images in order to know how information is shared out all over the bands. The results were satisfactory because information is spread all over the images bands and the algorithms used also have detected many interest points. Based on the results, a large database can be set up for a face recognition system building.**

Keywords*—Multispectral image; hybrid sensor; image mosaicing; image demosaicing; mosaic filter

I. Introduction

Nowadays, in almost every sector, security and attack problems have become a crucial challenge. Biometric imaging systems are appearing as a promising solution to increase levels of security. These biometric systems are mostly based on grayscale images, color images and spectral reflectance. But these systems still face tremendous difficulties when recognizing objects or faces.

In fact, conventional digital color cameras that generally operate in the visible spectrum seem to be limited in many situations where more information is needed and acquisition conditions are still difficult such as making acquisitions under a cloudy sky, while information beyond the visible range is required such as plants that emit in infra-red range, or when acquiring an image with more accuracy is needed, or when the calibration of the acquisition system is needed, or when making acquisitions in uncontrolled or dark environments is necessary, etc. Several studies have shown that images acquired in the visible spectrum present less information than those taken in Infra-red range [1], [2], [3], [4]. In addition Samuel ORTEGA et al. found that multispectral imaging technique able to obtain both spatial and spectral information within and beyond the human visual sensitivity, capture information regarding differents wavelengths [5]. To overcome some of these problems, MAMADOU Diarra et al, in their studies on multispectral images, have merged information from visible range and thermal infra-red to increase information in the image [6], [7], [8], [9], [10]. They also presented multispectral imaging and especially merging of information from the visible and infrared as a very promising alternative for image recognition. Moreover, Xingbo Wang et al. also found that for having more accuracy, it is necessary to make good choice of spectral characteristics of the camera's filters [11], [12], [4]. Their results show that the filter bandwidth had an influence on the accuracy of the reflectance estimation. However, multispectral imaging with cameras equipped with hybrid sensors, operating in the field of Near infra-red are much more efficient and can capture more information [1]. For example to verify that a fingerprint comes from a living finger and not a copy of that finger, it is obvious that the near infra-red range is the best fit since veins are visible through the skin in this area, Laura Rey-Barroso et al. introduced Near infra-red (NIR) multispectral imaging system to evaluate deeper skin layers thanks to higher penetration of photon at this wavelengths [13]. This hybrid system, integrated into a camera with dedicated hardware and software computations, allows a performance in real-time application with 30 fps. It also provides finer detail analysis in recognition systems.

In the context of the problems listed above, a camera equipped with a hybrid sensor has been proposed, in which an optimization of the spectral bands from 680 nm to 950 nm (NIR) has been set up as optimal bands [14]. This camera that captures images on eight bands allowed good resolution for images. These images were used in order to extract the characteristics for the recognition so that performances of recognition systems could be improved. Based on the results, a large database of images taken in the NIR can be set up.

In the following, the process consists of acquiring an image that will be mosaicked before being transmitted by the camera. Once we have the image from the camera, we proceed to the separation of the different spectral bands using binary masks (Fig. 2). After separation each spectral band contains only one spectral component. In order to get a complete image, these image bands have to be interpolated. This process is called demosaicking and it allows us to have complete image bands Fig. 3, Fig. 4 shows the entire acquisition process. After this last step, the interest points will be extracted in these image bands for tests.

II. Multispectral Hybrid Sensor

A. Hybrid Sensor

In this work a camera equipped with a hybrid sensor was used. This sensor was integrated into camera with a dedicated hardware unit, allowing the operation in real-time applications with 30 fps. In order to provide an optimal solution for the loss of spatial resolution inherent to MSFA, specific algorithms have been developed for multispectral demosaicking. The CMOS sensor is the physical element whose performance impacts on the quality of the final system. This sensor has been chosen respect to several criteria: $\surd$ Minimun pixel size is 5μm;

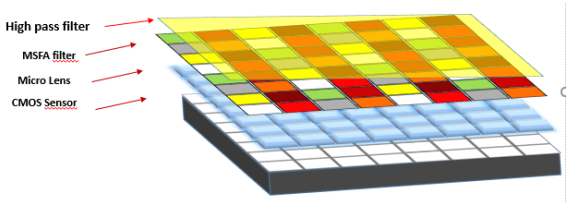
$\surd$ The CMOS sensor resolution should be high enough to compensate the loss related to the MSFA system;

$\surd$ The spectral sensitivity of the sensor must be extended to the near infra-red.

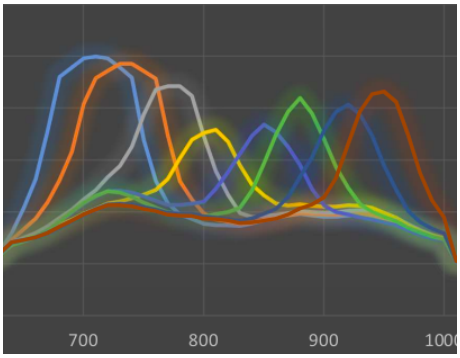
Taking into account the specifications above, our choice fell on the viimagic 9220H sensor. This sensor was provided by Grass Valley. Some modifications have been introduced in order to improve the final sensor.

The advantage of using CMOS sensors is that its manufacturing is much cheaper than CCD (Charged coupled Device) sensors. Furthermore, CMOS sensors consume less energy [15]. The ease of access to pixels available in CMOS sensors allows great flexibility for real-time data processing. All the above mentioned advantages bring about smaller systems, lower power consumption and lower manufacturing cost [15], [16]. As a result, we have chosen to use CMOS sensors rather than CCD sensors [17], [18]. The result of the mounted hybrid sensor and it's spectral response are presented in Fig. 1.

(a) Hybrid Sensor



(b) Hybrid Sensor Blocks



(c) Hybrid Sensor Spectral Response

Fig. 1. The Resulting Hybrid Sensor

B. Multispectral Images

A multispectral (MS) image is an image acquired by a sensor that operates in several spectral bands; it can be defined as an image where each pixel contains essentially information on the reflectance of the scene. It is represented by the matrix of pixels as follows:

$$M = \left(M_1, M_2, \cdots, \cdots, M_j\right)$$

Where M_j is the associated matrix of j^{th} band of image.

$$M_j = \begin{pmatrix} x^j_{11} & x^j_{11} & \dots & x^j_{1n} \\ x^j_{21} & x^j_{22} & \dots & x^j_{2n} \\ \dots & \dots & \dots & \dots \\ x^j_{m1} & x^j_{m2} & \dots & x^j_{mn} \end{pmatrix}$$

Let I be a MS image, a pixel of the image is noted $P(x, y)$, where x and y are the coordinates of the pixel P. Each pixel P is associated to a point $I(x, y, k)$ defined in a K-dimensional space (K being the number of component), and $I^k_{(x,y)}$, $k\epsilon\,\{1, 2, \dots, K\}$ represents the value of each component. Therefore, for a multispectral image one needs k components plans I^k, $k\epsilon\,\{1, 2, \dots, K\}$. In this study $K = 8$, called an 8-band multispectral image.

III. Material and Methods

A. Mosaic Filters

Mosaic Filters are filters presented as a matrix where each filter is associated with a specific spectrum. These filters make it possible to divide finely the spectrum and thus to differentiate the bands. In this work, a set of 8 filters based on the principle of Fabry-Perot has been used. Table I illustrates the response of each of the eight filters. The resulting distribution of MSFA (Multispectral Filter Array) moxel is indicated in Table II.

TABLE I. Filter Bands Responses

Bands	λ(nm)	$\delta\lambda$(nm)
P1	717	32
P2	751	32
P3	776	31
P4	810	31
P5	835	31
P6	870	30
P7	895	31
P8	930	31

TABLE II. Spatial Distribution of a Moxel

P1	P5	P2	P6
P7	P3	P8	P4
P2	P6	P1	P5
P8	P4	P7	P3

B. Mosaic Images

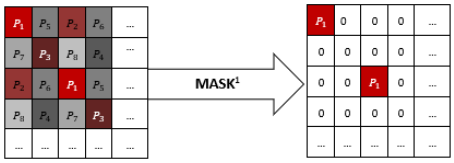
Image Mosaicing is a technique that allows building an image by superimposing successive images by registration [16], [19]. It can therefore be defined as the process of assembling different images of the same scene to form a single image [20]. The aim of mosaic creation is to visualize a large area on a single image under perspective projection. One of its applications is the construction of large aerial and satellite images of small photographs collections [21].

C. Strips Extraction

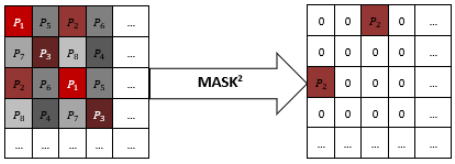
There are many algorithms used for strips extraction [20], [22]. In this work we have chosen to multiply the mosaic image by different binary masks $M^k_{(x,y)}$, $k \in \{1, 2, ..., K\}$ [20] which divides the mosaic image into $K = 8$ components. These masks have the value 1 at the positions where the pixel is available, and 0 at the other positions. Each component plan is obtained by multiplying the mosaic image term by term by the corresponding M^k mask.

By multiplying the mosaic image with each mask, we obtain 8 uncorrelated image plans (Fig. 2) on which only one spectral component is available. Each mask corresponds to an image plan.

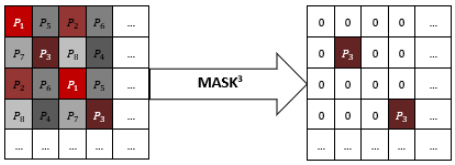
$$I'^k = I \odot M^k \quad (1)$$



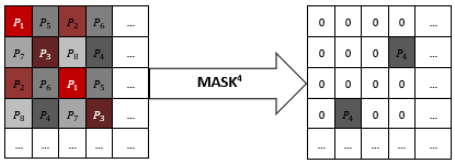
(a) Band1extract



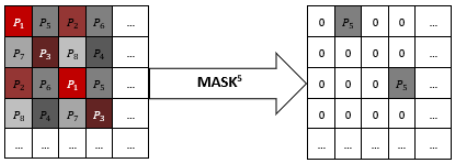
(b) Band2extract



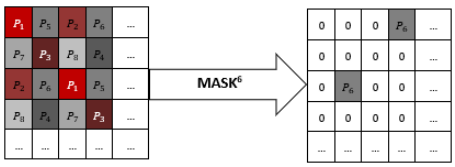
(c) Band3extract



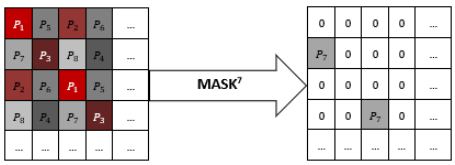
(d) Band4extract



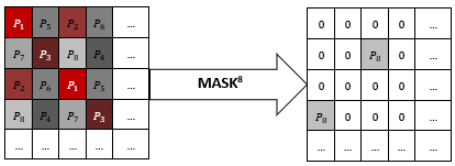
(e) Band5extract



(f) Band6extract



(g) Band7extract



(h) Band8extract

Fig. 2. Image Map after Applying Different Masks: After Applying Masks on Mosaic Image, Bands Obtained have only One Spectral Component

D. Multispectal Dematrixing by Bilinear Interpolation

After masks application, the resulting image plans contain only one spectral component. For complete image reconstruction, the missing pixels have to be interpolate. This process is called multispectral image demosaicing [23], [24], [18], [25], [3]. Bilinear interpolation [26], [16], [27], can be interpreted as a process of two linear interpolations, one in each direction. Linear interpolations can be made in several directions. $P(i,j)$ being the missing pixel at the position (i,j), we have:

- Diagonally:

$$P(i,j) = \frac{1}{4} \sum_{(m,n)=(-1,-1),(-1,1),(1,-1),(1,1)} p(i+m, j+n), \quad (2)$$

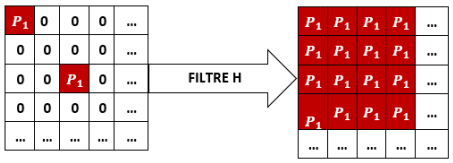
- Vertically:

$$P(i,j) = \frac{1}{2} \sum_{(m,n)=(-1,0),(1,0)} p(i+m, j+n), \quad (3)$$

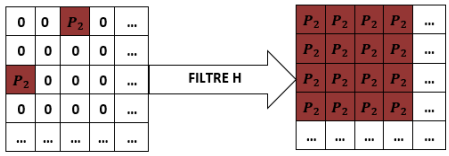
- Horizontally:

$$P(i,j) = \frac{1}{2} \sum_{(m,n)=(0,-1),(0,1)} p(i+m, j+n), \quad (4)$$

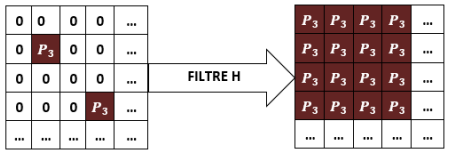
The interpolation or demosaicing of a mosaic image is a method that estimates the missing pixel on different (chromatic) channal of the mosaic image. Several algorithms have been designed for image demosaicing [28], [29], [30]. The method used, consists of applying convolution filter H on each band of the image obtained [20]. This filter is fixed so that the contribution of the neighbors in the pixel estimation of missing level in this pixel depends on the spatial distance separating the neighbor from the central pixel. Given that the pixels have the same structure, the same filter as shown in Mihoubi's work [20] is used. Interpolated bands have been shown in Fig. 3. The acquisition process is depicted in Fig. 4.



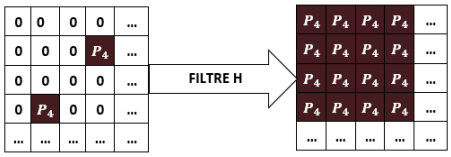
(a) Band1Interp



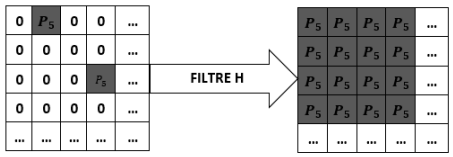
(b) Band2Interp



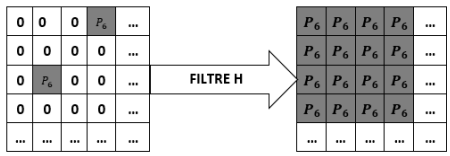
(c) Band3Interp



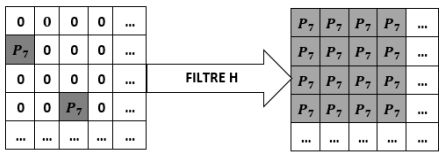
(d) Band4Interp



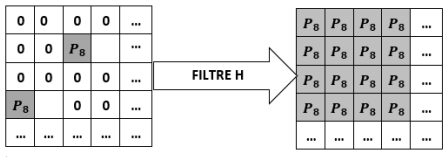
(e) Band5Interp



(f) Band6Interp



(g) Band7Interp



(h) Band8Interp

Fig. 3. Bands after Interpolation: The One Component Bands have been Interpolated using Filter H, to a Complete Image Bands

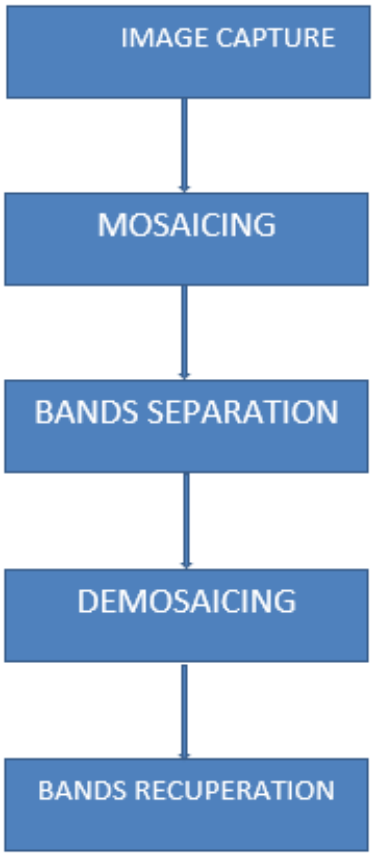


Fig. 4. Acquisition Process: The Process from Acquisition to the Last Band Recuperation

$$I^k = I'^k \odot H \tag{5}$$

I^k, is the interpolated image band.

$$H = \frac{1}{9}\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 4 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 4 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{pmatrix}$$

E. Point of Interest

A point of interest in an image is an area of pixel having remarkable properties often expressed by abrupt changes in intensity. They are regions of the image rich in terms of local information content and stable under affine transformations and illumination variations. In an image there must be few points whose local descriptors are similar [31].

IV. Extraction of Interest Points

The feature extraction methods are based on Scale Invariant Feature Transform (SIFT). The SIFT detector [32] is the best known of the detectors. This method combines a detector with a descriptor. SIFT's point of interest detection is based on a DoG (Difference of Gaussian), and has several versions. These algorithms are used in several contexts as multispectral imaging, face recognition under different criteria so that the performance of such a feature extraction kernel be able to extract the parameters [33]. It should be remembered that the SIFT method is based on the determinant of the Hessian matrix.

$$H(x,\delta) = \begin{pmatrix} L_{xx}(x,\delta) & L_{xy}(x,\delta) \\ L_{xy}(x,\delta) & L_{yy}(x,\delta) \end{pmatrix} \tag{6}$$

where $L_{xx}(x,\delta)$ is the convolution of the second order Differential of the Gaussian (DoG), which is the same for $L_{xy}(x,\delta)$ and $L_{yy}(x,\delta)$ to reduce the computation complexity of the determinant that uses the approximation of the wavelets of Haar.

$$H_{approx}(x,\delta) = \begin{pmatrix} D_{xx}(x,\delta) & D_{xy}(x,\delta) \\ D_{xy}(x,\delta) & D_{yy}(x,\delta) \end{pmatrix} \tag{7}$$

By using the expression of the integral image :

$$I_\Sigma(x) = \sum_{i=0}^{i\leq x}\sum_{j=0}^{j\leq y} I(i,j) \tag{8}$$

it can be deduced that:

$$det(H_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2 \tag{9}$$

V. Impact of the Work

The acquisition with hybrid sensors is made to measure the accuracy and the response of the resulting optical filters, which can ensure the accuracy and quality of the obtained multispectral images. These Multispectral images of the hybrid sensor can be less good, because of the demosaicking that compute the neighboring pixels which sometimes generate approximations. But the hybrid sensors are adequate for making snapshot acquisitions in real time application and it use in the case of this work for detecting faces in real time. The multispectral images from a filter wheel camera are very good quality [34], no approximation in the calculations, however, it is impossible to make the detection in real time. With this new camera, a multispectral images database will be set up. When the database contains enough images, a Deeplearning solution will be proposed in future work, as many research projects are moving towards this solution. In 2019 Shaukat Hayat et al. [35], proposed to use deep CNN-based features for Hand-Drawn sketch recognition via Transfer Learning Approach. Xiang Wang et al. introduced also method of privacy-preserving face recognition [36] where the convolutional neural network is used for face feature extraction. Moreover Bogdan BELEAN et al. [37] use CNN (Convolutional Neural Network) for images segmentation.

VI. Presentation of Detectors and Descriptors

Face or shape recognition techniques require some tools such as detectors and descriptors that are complementary tools of object recognition.

A. Detectors

Point-of-interest detection is a preliminary step in many computer vision processes. Detectors are used to isolate areas of interest in an image. For twenty years, several interest-point detectors have been developed. Schmid and Mohr compared the performance of several of these detectors. According to Schmid et al. [38], the most popular point-of-interest detector is the Harris detector [39]. The Harris corner detector was proposed by C. Harris and M. Stephens [40]. This easily detects the point of interest through a small window by moving this window in any direction. The Harris corner detection algorithm is performed by calculating the gradient of each pixel. Then, if the gradient values in the two directions are both large, the pixel is assumed to be a corner. Our experiences have been done using KAZE, Harris, ORB. KAZE, ORB, which are at the same time detectors and descriptors [41].

ORB (Oriented FAST and Rotated BRIEF) was introduced by Rublee et al. [42]. The Oriented Fast and Rotated Brief algorithm is based on the BRIEF keypoint descriptor and the FAST keypoint detector since both algorithms are computationally fast. It was presented in 2011 to provide a fast and efficient alternative to SIFT [43]. It is a variant of BRIEF to fill the lack of rotational invariance of it. The ORB method calculates a local orientation using an intensity centroid, which is defined as a weighted average of the pixel intensities in the local patch assumed not to coincide with the center of the entity.

The KAZE algorithm was developed in 2012 and it is in the public domain. The name comes from the Japanese word kaze which means wind and makes reference to the flow of air ruled by nonlinear processes on a large scale [44], [43]. For object recognition KAZE follows mainly the same steps as SIFT but with some differences in each step. KAZE algorithm [44], [45], instead of using DoG use AOS (Additive Operator Splitting) method and the Hessian matrix detector for blobs detection (DoH : Determinant of the Hessian) [43], [46].

$$L_{Hessian} = (\sigma)^2(L_{xx}L_{yy} - (L_{xy})^2) \tag{10}$$

where $L_{xx}(x, \delta)$ is the convolution of the second order Differential of the Gaussian (DoG), which is the same for $L_{xy}(x, \delta)$ and $L_{yy}(x, \delta)$.

B. Descriptors

After detecting points of interest, descriptors are used to describe them. They analyze neighborhood of each point to produce a characteristic vector of the interest point area. This vector is called the descriptor vector and in our work this vector describes 64 features. The description vector associated with a point of interest is a set of values extracted from the image in the local neighborhood of the position of the detected point [47]. This work have utilized the detectors and binary feature descriptors in Table III that provide high performance and compact data representation [38], [39].

TABLE III. SET OF DETECTORS AND DESCRIPTORS USED

Detectors	Descriptors
ORB	ORB
KAZE	BRISK
KAZE	FREAK
KAZE	KAZE
Harris	BRISK
Harris	FREAK
Harris	KAZE

BRISK (Binary Robust Invariant Scalable Keypoints) descriptor algorithm has been proposed by Leutenegger et al. [48]. In its detection, it uses the AGAST (Adaptive and Generic Corner Detection Based on the Accelerated Segment Test) [49] which is an improved variant of FAST [50]. FREAK (Fast Retina Keypoint) is a binary descriptor proposed by Alahi et al. [51]. Like BRISK (Binary Robust Invariant Scalable Keypoints), this descriptor uses a sampling model and a compensation method orientation. This is a variant of BRISK improved using a selection of pairs of templates. FREAK organizes sampling points analogous to the structure of the biological retina. For the description of the point of interest, the tools used are weighted Gaussians, the motif functioning as the retina and an orientation assignment is made for the description.

VII. RESULTS

In this part, the results of different stages of this work will literally be presented: mosaicing of the image, the decomposition into 8 bands and the points of interest tests.

A. Mosaicing of the Image and Decompositions into 8 Bands

The image obtained after mosaicing is represented in Fig. 5.

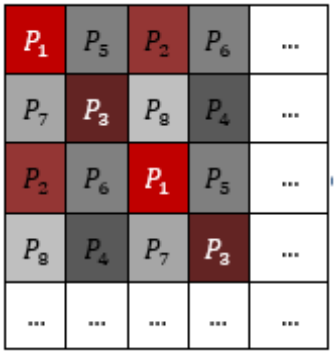
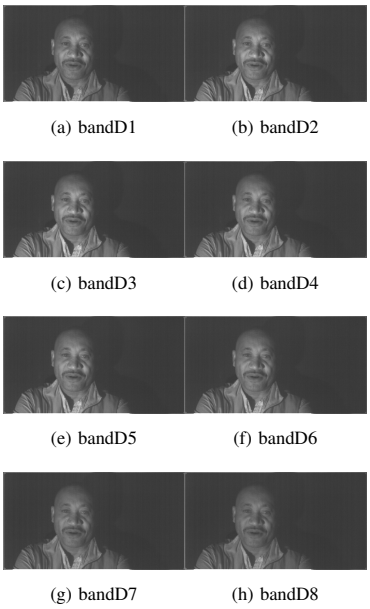


Fig. 5. Motif of Filter: The 4x4 Moxel Used

By applying different masks, mosaic image will be separated into 8 bands of images whose pixels contained a single color component, Fig. 2.

After the separation of the strips, these strips are demosaiced in order to attribute the rest of the color components to each pixel (Fig. 3). Fig. 6 representes a sample of final results on 8 bands after acquisition.



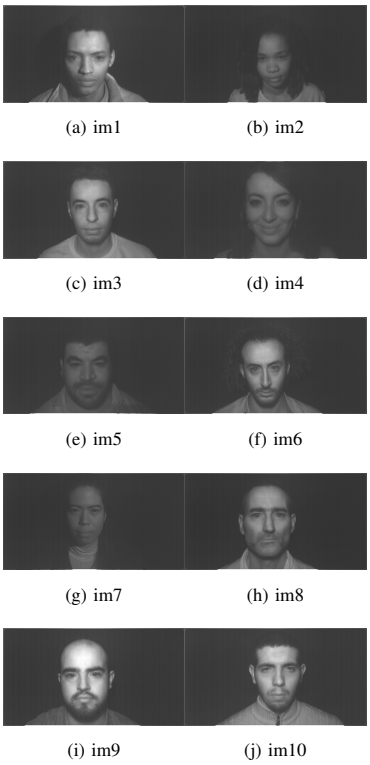
(a) bandD1 (b) bandD2

(c) bandD3 (d) bandD4

(e) bandD5 (f) bandD6

(g) bandD7 (h) bandD8

Fig. 6. Image on 8 Bands after Acquisition Process: Example of Bands Recuperated



(a) im1 (b) im2

(c) im3 (d) im4

(e) im5 (f) im6

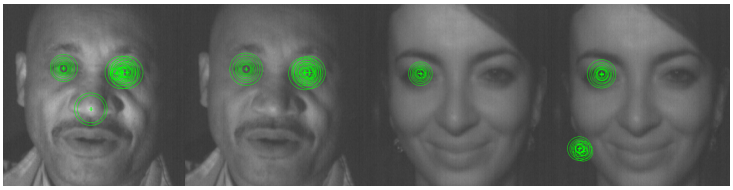
(g) im7 (h) im8

(i) im9 (j) im10

Fig. 7. Sample of Images of Database Used

B. Test and Images Used

The images data base is set up with images taken by a camera equipped with a hybrid sensor that detect and acquire faces in real time. This camera takes images on 8 bands and can be used in real time applications. Most of the time, for multispectral images, some bands contain less information than others. But the particularity of our camera is that the information is roughly spread over all the bands of image.The interest points are detected on all the 8 bands of images for all the algorithms mentioned above. Since the recognition is done on the face only, we used the algorithm of Viola Jones [52] to crop the face before the detection of those points of interest. This algorithm allows detecting only regions of interest. On the resulting image, different algorithms for the detection and the description of the points of interest has been applied.The tests have been done using Matlab v2020a with a sample of 30 images and the results are almost the same on each image. 10 images is used in this paper (Fig. 7). A sample of interest points by ORB/ORB has been shown in Fig. 8.



(a) (b) (c) (d)

Fig. 8. Example of Key Points on Bands: Key Points with ORB / ORB Detector / Descriptor

Each detected point is described in 64 elements that are unique even if the position or the place of image acquisition change. The tests and results have been presented in Table IV.

Given the results in Table IV, the KAZE algorithm detects more points of interest than others and The results show that the points of interest are slightly more concentrated on the first bands for all the algorithms other than Harris algorithm which detects more interest points on the last bands. But in general, the information spread over all eight bands if the acquisition process has been successful.

C. Entropy Test

Entropy in an image, makes it possible to measure the quantity of information contained in the image. In this work it allowed us to confirm that, information is spread over all the 8 bands. This entropy is computed by the formula below:

$$S = -\sum_{i=1}^{i \leq n} P_i \log(P_i) \tag{11}$$

Where P_i is the probability of each pixel occurrence.

Entropy tests have been done and the results are recorded in Table V and Fig. 9 represents the associated histogram.

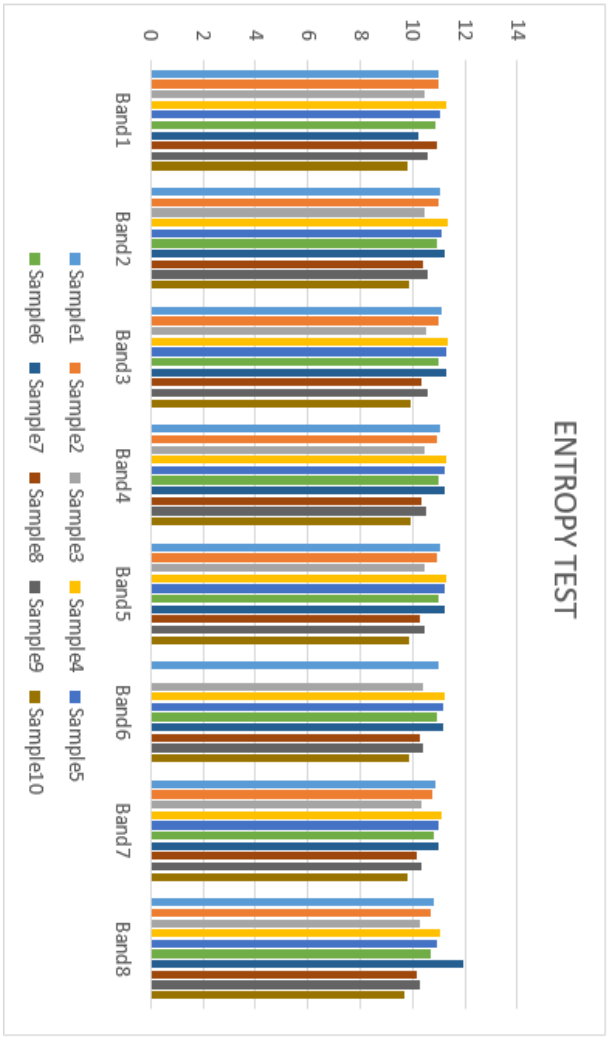


Fig. 9. Histogram of Entropy Test: Information Distributed over All the 8 Bands

VIII. Discussion

The results of Table IV allow to realize that when one used the pairs of detectors and descriptors Harris/FREAK, Harris/BRISK or Harris/KAZE, they did not detect enough interest points, especially on the first 6 bands. But on the seventh and eighth strips of some images, the number of interest points is quite enough. This phenomenon could be due to the lighting of the scene or the fact that this algorithm is not robust or suitable for these types of images. These algorithms have the worst performance in most cases in terms features detected. The algorithm ORB/ORB has correctly detected the points of interest on each band and each image. Therefore ORB/ORB demonstrates fairest precision with respect to the features, due to its performance one can say that it is better than the Harris/FREAK, Harris/BRISK and Harris/KAZE pairs which did not detect enough points and were also not stable for these types of images. For the pairs: KAZE/BRISK, KAZE/FREAK and KAZE/KAZE, the results were very satisfactory. The algorithms were performant at detecting high number of interest points. So, KAZE detector associate with BRISK, FREAK and KAZE descriptors are efficiency and robust enough for these multispectral images. By making a comparison between these three pairs, KAZE/BRISK, KAZE/FREAK and KAZE/KAZE, in general, it is KAZE/KAZE detector/descriptor pair which is better represented than the others in terms of features detection. This result confirm that of Shaharya et al. [53], [54], where KAZE in term of interest points outperform ORB. Furthermore, Ratsimbazafy et al. have shown in [55] that in terms of detection: SURF and KAZE are in a high category compared to ORB. By talking about stability, KAZE is more stable. But on the other hand, in terms of execution time, KAZE is not as efficient as ORB. Shuvo Kumar Paul et al. studying detector pairs [56] find that the KAZE and AKAZE pairs perform better than other pairs. This algorithm has detected several points of interest on each image and on each strip. This proves that the KAZE/KAZE pair would be suitable for these multispectral images from the camera equipped with a hybrid sensor and operating in the near infrared range. One can notice that the interest points were almost spread on all the bands. For confirmation, we compute the entropy tests. This entropy tests in the Table V should show how the information in each image is distributed. These entropy results showed that for each image the information is almost roughly distributed over all the 8 bands except the 7th and eighth bands which detects less points of interest than the others. However, the results of the entropy tests confirm the interest points results.

Based on these results, a large database of images taken on 8 bands with this camera which operates in the NIR can be set up.

IX. Conclusion

Security challenges of information systems keep increasing. Researchers have proposed different approaches and techniques. One of them is biometric imaging system. In recent years, studies have shown the limitation of this approache. This study focuses on multispectral (MS) imaging, primarly the use of the camera equipped with a hybrid sensor. This MS camera used in this work was built with a hybrid sensor, a Multispectrsl Filter Array (MSFA) mounted on a CMOS sensor that provided the best resolution for mosaic image, due to the small moxel used and due to the size of filter pitch $(5 \times 5\mu m2)$. This new camera system operates in the field of near infra-red in order to improve the process of object or image recognition. This study looked at the performance of this multispectral camera built in ImViA laboratory at the University of Bourgogne by extracting the points of interests on the bands of the multispectral images acquired by this camera. It also have been shown how to transform the obtained row images directly from the camera to a multispectral image through different steps namely: mosaicking, interpolation or demosaicing. Different descriptors have been used to extract interest points and the results were satisfactory. KAZE descriptor was the best and should be used to build recognition systems. However this project did not take place without difficulties: the filter based on the principle of Febray-Perrot is penalized by its secondary response; to compensate for the loss of sensitivity beyond 850 nm, a complex structure of moxel (6 x 6 pixels) simulation have been adopted, but this solution leads to none homogenous distribution. Ultimately,a regular distribution of the pixels (4x4) in the moxel is kept. The images are so contrasted that some lest robust algorithms are not able to extract interest points.

The Multispectral images of the hybrid sensor can be less good, because of the demosaicking that compute the neighboring pixels which sometimes generate approximations. But the hybrid sensors are adequate for making acquisitions and detecting faces in real time. Compared to the multispectral images from a filter wheel camera which are very good quality, no approximation in the calculations; however, it is impossible to make the detection in real time.

Future works will focus on building a large database of images acquired with this camera following by the usage of machine learning for recognition systems.

REFERENCES

[1] Wilder, Joseph and Phillips, P Jonathon and Jiang, Cunhong and Wiener, Stephen, Comparison of visible and infra-red imagery for face recognition, Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 182–187, 1996, IEEE

[2] Liu, Jun and Kumar, Ajay, Detecting presentation attacks from 3d face masks under multispectral imaging, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages=47–52, year=2018

[3] Balbhimrao, Lamb Anupama and Khambete, Madhuri, No-Reference Perceived Image Quality Algorithm for Demosaiced Images, INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 7, 1, 285–289, 2016, Citeseer

[4] He, Xin and Liu, Yajing and Ganesan, Kumar and Ahnood, Arman and Beckett, Paul and Eftekhari, Fatima and Smith, Dan and Uddin, Md Hemayet and Skafidas, Efstratios and Nirmalathas, Ampalavanapillai and others, A single sensor based multispectral imaging camera using a narrow spectral band color mosaic integrated on the monochrome CMOS image sensor, APL Photonics, 5, 4, 046104, 2020, AIP Publishing LLC

[5] Ortega, Samuel and Halicek, Martin and Fabelo, Himar and Callico, Gustavo M and Fei, Baowei, Hyperspectral and multispectral imaging in digital and computational pathology: a systematic review, Biomedical Optics Express, 11, 6, 3195–3233, 2020, Optical Society of America

[6] Diarra, Mamadou and Gouton, Pierre and Jérôme, Adou Kablan, Multispectral face recognition using hybrid feature, Electronic Imaging, 2017, 18, 200–203, 2017, Society for Imaging Science and Technology

[7] Li, Wei and Dong, Mingli and Lu, Naiguang and Lou, Xiaoping and Zhou, Wanyong, Multi-sensor face registration based on global and local structures, Applied Sciences, 9, 21, 4623, 2019, Multidisciplinary Digital Publishing Institute

[8] Zhang, Yongtao and Yin, Zhishuai and Nie, Linzhen and Huang, Song, Attention based multi-layer fusion of multispectral images for pedestrian detection, IEEE Access, 8, 165071–165084, 2020, IEEE

[9] Grifoni, Emanuela and Campanella, Beatrice and Legnaioli, Stefano and Lorenzetti, Giulia and Marras, Luciano and Pagnotta, Stefano and Palleschi, Vincenzo and Poggialini, Francesco and Salerno, Emanuele and Tonazzini, Anna, A new Infrared True-Color approach for visible-infrared multispectral image analysis, Journal on Computing and Cultural Heritage (JOCCH), 12, 2, 1–11, 2019, ACM New York, NY, USA

[10] Tariq Ahmad and Jinsong Wu and Imran Khan and Asif Rahim and Amjad Khan Human Action Recognition in Video Sequence using Logistic Regression by Features Fusion Approach based on CNN Features, International Journal of Advanced Computer Science and Applications, http://dx.doi.org/10.14569/IJACSA.2021.0121103, 2021, The Science and Information Organization, 12, 11,

[11] Wang, Xingbo and Thomas, Jean-Baptiste and Hardeberg, Jon Y and Gouton, Pierre, Multispectral imaging: narrow or wide band filters?, JAIC-Journal of the International Colour Association, 12, 2014

[12] Deng, Lei and Mao, Zhihui and Li, Xiaojuan and Hu, Zhuowei and Duan, Fuzhou and Yan, Yanan, UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras, ISPRS journal of photogrammetry and remote sensing, 146, 124–136, 2018, Elsevier

[13] Rey-Barroso, Laura and Burgos-Fernández, Francisco J and Delpueyo, Xana and Ares, Miguel and Royo, Santiago and Malvehy, Josep and Puig, Susana and Vilaseca, Meritxell, Visible and extended near-infrared multispectral imaging for skin cancer diagnosis, Sensors, 18, 5, 1441, 2018, Multidisciplinary Digital Publishing Institute

[14] Ansari, Keivan and Thomas, Jean-Baptiste and Gouton, Pierre, Spectral band Selection Using a Genetic Algorithm Based Wiener Filter Estimation Method for Reconstruction of Munsell Spectral Data, Electronic Imaging, 2017, 18, 190–193, 2017, Society for Imaging Science and Technology

[15] Hizem, Walid, Capteur intelligent pour la reconnaissance de visage 2009, Evry, Institut national des télécommunications

[16] Phelippeau, Harold, Méthodes et algorithmes de dématriçage et de filtrage du bruit pour la photographie numérique, 2009 Université Paris-Est

[17] Carlson, Bradley S, Comparison of modern CCD and CMOS image sensor technologies and systems for low resolution imaging, SENSORS, 2002 IEEE, 1, 171–176, 2002, IEEE

[18] Frommen, Thorsten, Adaptive Homogeneity-Directed Demosaicing Algorithm,
2007

[19] Capel, David, Image Mosaicing and super-resolution, Image Mosaicing and super-resolution, 47–79, 2004, Springer

[20] Mihoubi, Sofiane and Losson, Olivier and Mathon, Benjamin and Macaire, Ludovic, Multispectral demosaicing using intensity-based spectral correlation, 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), 461–466, 2015, IEEE

[21] Shum, Heung-Yeung and Szeliski, Richard, Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment, International Journal of Computer Vision, 36, 2, 101–130, 2000, Springer

[22] Miao, Lidan and Qi, Hairong and Snyder, Wesley E, A generic method for generating multispectral filter arrays, 2004 International Conference on Image Processing, 2004. ICIP'04., 5, 3343–3346, 2004, IEEE

[23] Cao, Hong and Kot, Alex C, Accurate detection of demosaicing regularity for digital image forensics, IEEE Transactions on Information Forensics and Security, 4, 4, 899–910, 2009, IEEE

[24] Hirakawa, Keigo and Parks, Thomas W, Adaptive homogeneity-directed demosaicing algorithm, IEEE Transactions on Image Processing, 14, 3, 360–369, 2005, IEEE

[25] Baone, Gaurav A and Qi, Hairong, Demosaicking methods for multispectral cameras using mosaic focal plane array technology, Spectral Imaging: Eighth International Symposium on Multispectral Color Science, 6062, 60620A, 2006, International Society for Optics and Photonics

[26] Alleysson, David and Susstrunk, Sabine and Hérault, Jeanny, Linear demosaicing inspired by the human visual system, IEEE Transactions on Image Processing, 14, 4, 439–449, 2005, IEEE

[27] Longere, Philippe and Zhang, Xuemei and Delahunt, Peter B and Brainard, David H, Perceptual assessment of demosaicing algorithm performance, Proceedings of the IEEE, 90, 1, 123–132, 2002, IEEE

[28] Brauers, Johannes and Aach, Til, A color filter array based multispectral camera, 12. Workshop Farbbildverarbeitung, 2006, Ilmenau

[29] Chang, Edward and Cheung, Shiufun and Pan, Davis Y, Color filter array recovery using a threshold-based variable number of gradients, Sensors, Cameras, and Applications for Digital Photography, 3650, 36–43, 1999, International Society for Optics and Photonics

[30] Li, Xin and Orchard, Michael T, New edge-directed interpolation, IEEE transactions on image processing, 10, 10, 1521–1527, 2001, IEEE

[31] Calonder, Michael and Lepetit, Vincent and Strecha, Christoph and Fua, Pascal, Brief: Binary robust independent elementary features, European conference on computer vision, 778–792, 2010, Springer

[32] Mikolajczyk, Krystian and Schmid, Cordelia, A performance evaluation of local descriptors, IEEE transactions on pattern analysis and machine intelligence, 27, 10, 1615–1630, 2005, IEEE

[33] Albatal, Rami and Mulhem, Philippe and Chiaramella, Yves, Phrases Visuelles pour l'annotation automatique d'images., CORIA, 10, 3–18, 2010

[34] Johannes Brauers and Nils Schulte and Til Aach, Multispectral Filter-Wheel Cameras: Geometric Distortion Model and Compensation Algo-

rithms, IEEE Transactions on Image Processing, IEEE, 2008, Dec, 17, 2368–2380, 12,

[35] Hayat, Shaukat and She, Kun and Yu, Yao and Mateen, Muhammad, Deep cnn-based features for hand-drawn sketch recognition via transfer learning approach, Editorial Preface From the Desk of Managing Editor. . . , 10, 9, 2019

[36] Wang, Xiang and Xue, Heyu and Liu, Xuefeng and Pei, Qingqi, A privacy-preserving edge computation-based face verification system for user authentication, IEEE Access, 7, 14186–14197, 2019, IEEE

[37] Belean, Bogdan, Active Contours Driven by Cellular Neural Networks for Image Segmentation in Biomedical Applications, STUDIES IN INFORMATICS AND CONTROL, 30, 3, 109–119, 2021, NATL INST R&D INFORMATICS-ICI PUBL DEPT, 8-10 AVERESCU BLVD, SECTOR 1 . . .

[38] Schmid, Cordelia and Mohr, Roger and Bauckhage, Christian, Comparing and evaluating interest points, Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), 230–235, 1998, IEEE

[39] Schmid, Cordelia and Mohr, Roger and Bauckhage, Christian, Evaluation of interest point detectors, International Journal of computer vision, 37, 2, 151–172, 2000, Springer

[40] Mikolajczyk, Krystian and Schmid, Cordelia, Scale & affine invariant interest point detectors, International journal of computer vision, 60, 1, 63–86, 2004, Springer

[41] Cowan, Bruce and Imanberdiyev, Nursultan and Fu, Changhong and Dong, Yiqun and Kayacan, Erdal, A performance evaluation of detectors and descriptors for UAV visual tracking, 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), 1–6, 2016, IEEE

[42] Rublee, Ethan and Rabaud, Vincent and Konolige, Kurt and Bradski, Gary, ORB: An efficient alternative to SIFT or SURF, 2011 International conference on computer vision, 2564–2571, 2011, Ieee

[43] Andersson, Oskar and Reyna Marquez, Steffany, A comparison of object detection algorithms using unmanipulated testing images: Comparing SIFT, KAZE, AKAZE and ORB, 2016

[44] Alcantarilla, Pablo Fernández and Bartoli, Adrien and Davison, Andrew J, KAZE features, European Conference on Computer Vision, 214–227, 2012, Springer

[45] Noble, Frazer K, Comparison of OpenCV's feature detectors and feature matchers, 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP), 1–6, 2016, IEEE

[46] Ze-Ping, Cai and De-Gui, Xiao, Feature matching algorithm based on KAZE and fast approximate nearest neighbor search, 3rd International Conference on Computer Science and Service System, 2014, Atlantis Press

[47] Harris, Christopher G and Stephens, Mike and others, A combined corner and edge detector., Alvey vision conference, 15, 50, 10–5244, 1988, Citeseer

[48] Leutenegger, Stefan and Chli, Margarita and Siegwart, Roland Y, BRISK: Binary robust invariant scalable keypoints, 2011 International conference on computer vision, 2548–2555, 2011, Ieee

[49] Mair, Elmar and Hager, Gregory D and Burschka, Darius and Suppa, Michael and Hirzinger, Gerhard, Adaptive and generic corner detection based on the accelerated segment test, European conference on Computer vision, 183–196, 2010, Springer

[50] Rosten, Edward and Porter, Reid and Drummond, Tom, Faster and better: A machine learning approach to corner detection, IEEE transactions on pattern analysis and machine intelligence, 32, 1, 105–119, 2008, IEEE

[51] Alahi, Alexandre and Ortiz, Raphael and Vandergheynst, Pierre, Freak: Fast retina keypoint, 2012 IEEE Conference on Computer Vision and Pattern Recognition, 510–517, 2012, Ieee

[52] Jones, Michael and Viola, Paul, Fast multi-view face detection, Mitsubishi Electric Research Lab TR-20003-96, 3, 14, 2, 2003

[53] Tareen, Shaharyar Ahmed Khan and Saleem, Zahra, A comparative analysis of sift, surf, kaze, akaze, orb, and brisk, 2018 International conference on computing, mathematics and engineering technologies (iCoMET), 1–10, 2018, IEEE

[54] Ramkumar, B and Laber, Rob and Bojinov, Hristo and Hegde, Ravi Sadananda, GPU acceleration of the KAZE image feature extraction algorithm, ournal of Real-Time Image Processing, 17, 5, 1169–1182, 2020, Springer

[55] Ratsimbazafy, TH and Randriamitantsoa, PA Comparaison de performances des détecteurs KAZE, ORB, SURF,

[56] Paul, Shuvo Kumar and Hoseini, Pourya and Nicolescu, Mircea and Nicolescu, Monica, Performance Analysis of Keypoint Detectors and Binary Descriptors Under Varying Degrees of Photometric and Geometric Transformations, arXiv preprint arXiv:2012.04135, 2020

TABLE IV. KEY POINTS EXTRACTION WITH DIFFERENT METHODS

Images	Bands	ORB/ ORB	KAZE/ BRISK	KAZE/ FREAK	KAZE/ KAZE	Harris/ BRISK	Harris/ FREAK	Harris/ KAZE
Sample1	Band1	49	287	336	389	9	9	9
	Band2	44	344	408	480	4	4	4
	Band3	37	324	387	466	3	3	3
	Band4	37	291	355	425	5	5	5
	Band5	30	250	302	374	2	2	2
	Band6	27	227	282	343	3	3	3
	Band7	24	156	195	244	3	3	3
	Band8	18	124	156	204	2	2	2
Sample2	Band1	37	287	334	357	4	4	4
	Band2	34	251	300	335	3	3	3
	Band3	29	188	229	261	3	3	3
	Band4	25	141	172	202	7	7	7
	Band5	21	140	165	189	3	3	3
	Band6	24	102	133	151	5	5	5
	Band7	26	67	85	104	4	4	4
	Band8	17	46	62	75	3	3	3
Sample3	Band1	144	169	180	181	20	20	20
	Band2	139	178	191	195	24	24	24
	Band3	144	144	156	158	23	23	23
	Band4	145	128	134	136	25	25	25
	Band5	152	120	129	132	21	21	21
	Band6	126	99	106	107	19	19	19
	Band7	117	72	76	76	20	20	20
	Band8	95	55	58	58	15	15	15
Sample4	Band1	32	200	233	292	2	2	2
	Band2	30	201	248	310	7	7	7
	Band3	26	176	222	297	11	11	15
	Band4	20	151	185	262	2	2	2
	Band5	21	123	159	226	60	60	69
	Band6	13	113	143	197	2	2	2
	Band7	9	92	116	157	517	517	561
	Band8	8	60	78	116	106	106	121
Sample5	Band1	27	117	128	174	2	2	2
	Band2	24	132	172	230	2	2	2
	Band3	19	121	162	227	5	5	5
	Band4	20	111	142	209	8	8	8
	Band5	10	97	127	179	1	1	1
	Band6	11	77	98	152	341	341	361
	Band7	5	51	66	107	61	61	65
	Band8	5	43	50	80	172	172	192
Sample6	Band1	21	203	259	289	5	5	5
	Band2	21	217	285	330	3	3	3
	Band3	18	170	230	272	3	3	4
	Band4	23	151	211	246	18	18	18
	Band5	12	128	173	220	35	35	36
	Band6	12	114	160	189	7	7	7
	Band7	11	65	87	113	215	215	238
	Band8	11	42	68	85	253	253	270
Sample7	Band1	37	292	315	368	5	5	5
	Band2	40	308	345	402	5	5	5
	Band3	31	236	259	308	37	37	41
	Band4	31	190	203	239	4	4	4
	Band5	24	153	168	204	351	351	371
	Band6	23	133	144	180	9	9	9
	Band7	12	93	101	127	177	177	189
	Band8	11	63	69	86	278	278	293
Sample8	Band1	11	68	86	92	1	1	1
	Band2	11	69	86	94	2	2	2
	Band3	15	57	71	76	1	1	1
	Band4	9	45	56	59	1	1	1
	Band5	11	40	49	52	1	1	1
	Band6	12	33	41	44	2	2	2
	Band7	8	17	19	21	8	8	8
	Band8	6	16	18	18	1	1	1
Sample9	Band1	17	39	50	63	2	2	2
	Band2	6	25	40	55	4	4	4
	Band3	12	18	29	38	2	2	2
	Band4	13	16	20	26	3	3	3
	Band5	11	9	14	19	2	2	2
	Band6	5	5	11	17	3	3	3
	Band7	7	4	4	7	2	2	2
	Band8	7	4	4	6	11	11	11
Sample10	Band1	5	3	3	3	1	1	1
	Band2	7	3	3	3	3	3	3
	Band3	5	3	3	4	1	1	2
	Band4	3	2	4	2	15	15	17
	Band5	4	2	2	2	22	22	26
	Band6	3	2	2	2	45	45	49
	Band7	0	2	2	2	596	596	680
	Band8	0	2	2	2	273	273	295

TABLE V. ENTROPY TEST

Bands	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
Band1	10.9735	10.9679	10.4355	11.2693	11.0272	10.8871	10.2259	10.9361	10.5752	9.8139
Band2	11.0420	10.9775	10.4639	11.3286	11.1261	10.9460	11.2483	10.3736	10.5881	9.8807
Band3	11.1117	10.9752	10.4896	11.3205	11.2625	11.0012	11.2764	10.3494	10.5584	9.9089
Band4	11.0632	10.9234	10.4486	11.2925	11.2444	10.9879	11.2326	10.3120	10.4947	9.9028
Band5	11.0739	11.9339	10.4635	11.2739	11.2266	10.9748	11.2172	10.3028	10.4727	9.8868
Band6	11.0101	10.8779	10.4193	11.2393	11.1768	10.9426	11.1633	10.2719	10.4235	9.8687
Band7	10.8807	10.7572	10.3200	11.1083	11.0038	10.7959	11.0099	10.1796	10.3231	9.7819
Band8	10.8172	11.6894	10.2670	11.0272	10.9420	10.7253	11.9361	10.1363	10.2562	9.7181

The Effectiveness of CATA Software in Exploring the Significance of Modal Verbs in Large Data Texts

Ayman F. Khafaga
College of Science & Humanities, Prince Sattam bin Abdulaziz University, Saudi Arabia
Faculty of Arts &Humanities, Suez Canal University, Egypt

***Abstract*—This paper investigates the effectiveness of using and applying CATA (Computer-Aided Text Analysis) software in exploring the extent to which particular modals are significant in communicating the ideological and thematic messages of literary discourse. More specifically, the paper attempts to test the hypothesis that CATA software, including FDA (Frequency Distribution Analysis), KWICK (Key Word in Context), CA (Content Analysis), and TDA (Thematic Distribution Analysis) are effectively helpful in the linguistic and ideological analysis of modals in literary texts. To this end, the paper uses the frequency distribution analysis (FDA) and applies it to Edward Bond's *Lear* as a sample representing literary texts. Two modal verbs were selected to be computationally analyzed by means of the frequency distribution analysis in order to decode the different ideologies they carry in the discourse of the selected play. These are *will* and *must*. These modal verbs were computationally displayed within their contextual, total and indicative occurrences in the play under investigation to demonstrate the way they convey particular ideologies. Findings revealed that CATA software represented in its variable of FDA is highly contributive to communicating ideologies in the play under investigation. The paper further demonstrated two findings: first, via CATA software, analysts can easily arrive at the ideological significance of the various classes of words, including modal verbs that are used in literary texts; and, second, the analysis showed that only a few occurrences out of the total number of frequencies of the modal verbs at hand are indicative in conveying the hidden ideologies of their users.**

***Keywords*—CATA software; frequency distribution analysis; ideologies; modal verbs; Bond's Lear**

I. Introduction

It is perspicuously evident that the application of computer software contributes significantly to the linguistic and stylistic study of literary and fictional texts, particularly to decode the different themes and ideologies pertaining to this type of texts [1], [2] and [3]. These computational software offer analysts and researchers the ability to arrive at authentic, reliable, and credible results in an accurate and precise way more than any other analytical tools that would be conducted without the interference of computer [4]. The current study scrutinizes to reveal the various ideologies pertinent to the usage of the modal verbs in one of the literary texts represented in Edward Bond's *Lear*. Beyond the choice of the play under consideration, the rationale lies in the fact that the majority of fictional texts, particularly the literary one abound in huge number of modal verbs that are usually utilized for particular grammatical purposes, including the expression of obligation, possibility, desirability, certitude, etc. [5]. This paper, therefore, offers a linguistic investigation of the importance of modal verbs in the play at hand by using two analytical dimensions. First, by using and applying the frequency distribution analysis (FDA), which will be activated by the program of concordance in order to display the different occurrences pertaining to each modal verb under investigation (i.e., will and must). Second, to reveal the ideological importance these modal verbs are employed to communicate in Edward Bond's *Lear*. Crucially, the current article scrutinizes to test the hypothesis that the incorporation of the latest developments of applying technology into the textual and contextual analysis of texts contributes significantly to the general understanding of texts, specifically, large data texts manifested in the context of this paper in the conversational literary genre.

In fact, with the unrelenting development of technology, computer software has become increasingly important in various types of corpus linguistics studies, where it is used to extract both theoretical and empirical conclusions toward textual analysis in general and to the field of linguistics specifically [6], [7], [8], and [9]. The findings of these studies indicated that computer software provides essential support and facilitation for a wide-ranging and improved analytical environment, in which analysts and stylisticians are able to conduct analysis in an efficient and effective way. This is conducted by providing ample, credible, and adequate results. Incorporating computational techniques in corpus linguistics studies not only facilitates the entire text analysis process, but also highlights the inclusion of technology as well as other social and human sciences into the research pertinent to corpus linguistics, stylistics, pragmatics and the various scopes of discourse studies [10].

In terms of its theoretical framework, the paper is grounded on two analytical frameworks: the first constitutes Fairclough's [11] approach to critical discourse analysis, in which he investigates various grammatical concepts relevant to the study of ideology in discourse, either written discourse or spoken discourse. Fairclough's perspective to the analysis of the grammatical concepts focuses, among other concepts, on the use of modal verbs as conduits of particular ideological and discursive purposes. The second dimension to be used here is a computer-aided text analysis (CATA); this digital tool will analytically be enabled, as alluded before, by a frequency distribution analysis (FDA), through which the two modal verbs at hand will undergo a digital analysis that serves to arrive at the frequencies of each modal verb in text as well as its contextual environment in discourse. Significantly, core

concern beyond using both CDA and FDA in the analysis of Bond's *Lear* [12] ultimately functions to explore the ideologies conveyed by the modal verbs at hand, and to highlight analytically the integration of modern technology and critical discourse studies. This is conducted by demonstrating the significant part modal verbs have in communicating various ideologies in discourse through showing the significant and insignificant occurrences of the modals under investigation.

The application of CATA via the frequency distribution analysis to linguistically investigate modal verbs as conduits of ideology in discourse mirrors the extent to which modal verbs in general and the modals under investigation in particular have linguistic and ideological weight in communicating meanings among conversationalists, either at the fictional level of discourse, as is the case in the current study, or in reality, that is, in everyday occurring conversations [13]. Crucially, large data texts are a fertile field in which computer software can be applied to uncover the hidden meanings and ideologies encoded in these texts [14-15]. The assumption that discourse is ideologically-loaded makes us proclaim that any type of discourse carry some sort of ideological importance. That is, any text is supposed to have an ideological message. These ideological messages can be decoded by means of various analytical tools. Among these tools are CDA and CATA, which are adopted in the analysis of this article [16]. Regardless of the fact that communication in literary texts is totally fictional, it remains a fact that these fictional communications reflect what is happening in the real world [17]. This, of course, shows the reason why *Lear* as a representative of large data texts is selected for the analysis in the current article.

A. Research Questions

Three research questions are attempted to be answered in this article as follows:

1) How does an FDA contribute to the analysis of modal verbs in fictional texts?

2) What are the various ideologies conveyed by the modal verbs in Bond's *Lear*?

3) To what extent does Key Word in Context (KEWIC) variable contribute to the intelligibility of the weight of specific modal verbs within particular contexts in literary discourse?

B. Research Objectives

Three research objectives are attempted to be achieved in this paper. These are as follows:

1) To show the way CATA contributes to the studies of ideologies in discourse, particularly in relation to the modal verbs investigated in the current article.

2) To highlight the harmonizing connection between CDA and CATA.

3) To explore the different ideologies modal verbs communicate in the play under investigation.

In the remaining part of this article, the paper will present the theoretical preliminaries and the literature review in Section II. The study also provides the method adopted in its analytical part, as well as the analytical procedures in Section III. In Section IV, the paper offers the analysis of the selected data. The discussion of the obtained results will be the focus of Section V, whereas the Section VI is dedicated to the conclusion and recommends some ideas for further research.

II. Literature Review

A. Computer-Aided Text Analysis

According to [18], the use of computational software in the textual and contextual analysis of texts is highly contributive to deciphering the hidden ideologies and the various propositional meanings pertaining to these texts. He also emphasizes that these software can be employed within the framework of computer assisted language learning (CALL) to facilitate the process of learning and teaching. It is also obvious that computer software contribute significantly to the linguistic and stylistic investigation of large data texts, particularly the fictional and literary ones. Such computational application of various programs and software makes it easy for linguists and discourse analysts to arrive at specific ideologies and meanings at the various levels of the linguistic analysis; that is, the level of the word, the level of the phrase, the level of the sentence, and the level of the utterance. This, in turn, accentuates the importance of the work of computer in the analysis of the different discourse genres, as it helps arrive at concise and credible results that are supposed to be thorny if the linguistic investigation is conducted manually, that is, without the computational work [19]. Such contributive role of computational analysis is not only manifested in the studies within the scope of linguistics, but it also exceeds this analytical focus to cover other fields, either in linguistics or in other social disciplines. Thus, computer software can be applied to studies in the fields, such as pragmatics, semantics, stylistics, etc [20]. This computational perspective towards the analysis of the various types of texts ends any case in which the reliability, authenticity and credibility of analysis that is conducted without computer. This is because computer serves to arrive at very authentic and concise results for any text [21].

Importantly, computer-assisted text analysis provides a wide range of analytical methods and possibilities that are valuable in the stylistic, semantic and pragmatic investigation of gigantic data texts, particularly fictional writings. The Frequency Distribution Analysis (FDA) is one of these analytical options, and it solely allows you to see how many times a searched item appears in a text. According to [22], frequency analysis allows analysts to acquire a rough notion of the textual nature of specific lexis in a text. This also aids in steering the analytical wheel toward a significant precedence of one occurrence over another, which is made possible by CATA's second variable, the Key Word in Context variable, which is computationally enabled (KWIC). The context in which a searched phrase appears is indicated by the KWIC variable. To put it another way, KWIC explains the context of the searched items, which aids in determining the ideological value of words and/or sentences [23]. Content Analysis is another analytical option provided by CATA (CA). For [24], content analysis is used to divide words into classes based on their semantic characteristics. Researchers and text analysts

use content analysis in conjunction with other CATA factors to determine the thematic and ideological relevance of certain words in texts.

In the context of this article, CATA is enabled to give the analytical alternatives indicated above by the Concordance program. Concordance is a computer application that allows analysts and users to collect, access, classify, and analyze various sorts of texts, especially those that contain enormous amounts of data [25-26]. Concordance can thereby retrieve all instances of a searched lexis in a text, display the contextual context of any word, and categorize all words based on their semantic meaning [27]. The ideologies represented by the modal verbs in the selected data will be revealed by deriving the frequency distribution of the modal verbs under consideration, which will be supplemented by both the use of KWIC and content analysis to reveal the ideologies represented by the modal verbs at hand.

B. Critical Discourse Analysis

Fairclough [11] presents four sets of items for the linguistic study of function words in texts and speech, based on his research into the role of modal verbs and their ideological relevance in discourse studies. The first set demands a study of textual experiential values, which necessitates a modality analysis. The second group focuses on the grammatical elements of texts, such as the style of phrase used (declarative, interrogative, or imperative); the type of modality utilized (truth, obligation, or possibility modality); and the type of pronouns used in discourse. The final group of items focuses on the expressive value of grammatical features like expressive modality. The fourth set examines the many sorts of sentences employed, such as simple, compound, and complicated, as well as the relationships between the various sentence forms. Fairclough's collections of items are rich in grammatical features that contribute to the production of ideologically charged discourse. Specifically, in terms of studies that pertinent to stylistics, ideology, pragmatics, discourse studies, and power relations [28], [29], and [30], modality is used discursively to express, develop, and retain agency. The reason for this, according to these studies, is that agency is inextricably linked to conceptions of power and dominance, and it is intricate to find any type of ideological discourse that does not present as its core concern notions and themes of power, control, persuasion, and manipulation. Modal verbs, in particular, can function as ideology carriers in language under certain settings.

C. Modal Verbs

These modal verbs are categorized into relational and expressive modality based on Fairclough's premise about modal verbs [11]. According to him, such cataloging is founded on the speaker's authority over his or her addressees. The focus in relational modality is on both the authority and discourse access of the speaker in relation to the various semantic propositions communicated by language. In expressive modality, on the other hand, the focus is on the authority of the speaker with regard to the degree of truthfulness concerning the content of discourses. Modality, he claims, can be articulated linguistically not only through modal verbs, but also via other devices of grammar, such as adverbs and tense. According to Fowler [31], there are four sorts of modality: truth, obligation, permission, and desirability. Modal auxiliaries like 'will' and adverbs of certitude like 'surely' can be used to communicate truth modality. This modality demonstrates that the speaker's assumption is correct. To communicate a high level of certainty, truth modals are utilized. Some modal auxiliaries, such as 'must, ‘should,' or 'ought to,' can be used to create an obligation modality. The obligation mode focuses on the perceptual attitudes of discourse participants towards the implementation of the semantic proposition pertinent to the speaker. The modal auxiliaries, such as 'can,' present the permission modality, which is employed to provide discourse participants the ability to communicate the language function of permission. The speaker's status of accepting or denying what is expressed by his/her offer is clarified by the desirability modality. In light of this paper, only two types of modal verbs will be discussed: truth modality (will) and obligation modality (must).

D. Related Studies

When looking at prior research on the usage and deployment of computer software in general, and frequency distribution analysis in particular, it is clear that these programs are quite useful for analyzing massive data sets like fictional novels. [32], for example, looked into the semantics, rhythm, and tempo of narrative storytelling using data mining. They determined how much input data might influence the final perception of fictitious literature and concluded that the process of data mining, which is conducted through visualization, can in turn mimic the semantic classification and thematic clustering carried by fictional texts.

Another study conducted by [33] looked into the effectiveness of concordance in the analysis of fictitious discourse. This research showed that concordance may be applied to massive data texts to produce legitimate and believable results that help with text comprehension. Both the FDA and the KWIC were the two analytical variables presented in this study, which were generated using concordance. The study indicated that utilizing and applying concordance to the investigation of literary texts aids in achieving a high level of intelligibility at the level of representation of various themes and ideologies, as well as determining the writer's intended meaning.

A further study presented by [34], in which they demonstrated the importance of using new technological software in developing reliable translation versions in the field of translation studies. The study advised using computer software into the teaching and learning of university translation courses in Saudi Arabia's various academic institutions. Furthermore, [35] conducted a study on the impact of CALL software on the academic competence and performance of Saudi university students, who are studying English as a foreign language. The use of CALL to EFL settings has a good impact on EFL students' learning results, according to this study. The main objective of this study is to see how effective the two computer apps SnagitTM and Screencast are at helping people learn to read. The study found that using the two computer software programs helps pupils enhance their academic performance by cultivating

language abilities that are important for learning to read. The study also found that incorporating technology into EFL classes helps students strengthen not only their linguistic skills, but also their communicative abilities. The study concluded by proposing the use of CALL software in the context of Saudi universities and in the course designation process at Saudi institutions for various EFL courses.

Within the scope of legal studies, [36] investigated the extent to which concordance helps analyze the linguistics of the opening statements by revealing the numerous pragmatic meanings and concealed ideologies that lies beyond the mere semantic propositions that are expressed at the surface level of the semantics of texts. The study further accentuates the effective role computational analysis provides to uncover the mystifying meanings and the pragmatic purposes attempted to be clarified on the part of writers and/or speakers. Significantly, this study states that these ideological activities within textual and contextual analyses can be revealed and conducted via the repeated usage of particular lexical elements in the investigated texts. The research stated above demonstrate the usefulness and contribution of computer software in linguistic studies, whether on fictional texts or outside the realm of fiction, such as in EFL and courtroom contexts. The usage of CATA to discover further and new meanings in texts, particularly those that are produced by function words in large data and fictional texts is expected to be expanded in this study so as to offer fertile insights into the contributive way CATA software are employed to explore the various meanings pertaining to this type of texts.

III. Methodology

A. Data Collection and Description

Edward Bond's Lear is included in this study's corpus. The play is divided into four acts, each of which contains eighteen scenes that make up the entire production of the dramatic piece. The reason for choosing this play in particular is that it contains a substantial amount of grammatical elements that are useful in transmitting various views, ranging from persuasive to manipulative. This is demonstrated by the frequency analysis included in this article, which demonstrates an ideological weight for such grammatical characteristics; they are not used haphazardly in the play's dramatic dialogue, but rather serve as ideological containers. Bond's Lear, in particular, has distinctive discourse features that allow an investigation of the ideologies encoded in the play at hand, particularly in terms of the deft use of modality to transmit beliefs and communicate ideological meanings.

As for the play itself, it is a rewriting of Shakespearean masterpiece King Lear. The play narrates the story of a king who has two daughters and, like Shakespearean king, he decides to surrender his kingdom to them. The two daughters betray their old father, dismiss him outside their homes, and cause him to madness. The play shows the extent to which political ideologies are communicated in discourse. It also sheds light on the way language is employed not only to persuade, but also to manipulate. In light of the current article, it will analytically be shown how modal verbs are utilized to convey such persuasive and/or manipulative ideologies. This, as mentioned before, will be conducted by the application of CATA software manifested in an FDA.

B. Analytical Procedures

The analytical procedures adopted in this study constitute three stages. Three CATA variables were used in each stage: frequency distribution analysis (FDA), key word in context (KWIC), and content analysis (CA). The first step was to prepare the text of the chosen play by electronically uploading it so that it could be analyzed. This stage gave an overview of how the play's discursive tone is conveyed through the characters' conversational turns. The modal verbs (will and must) were electronically illuminated in the second stage to mark their appearance in the play. This was accomplished by applying an FDA to the entire text of the play at hand, and tracking the various occurrences of each searched token. The final stage consisted of an interpretive exercise in which all of the highlighted elements were investigated in terms of their likely occurrences in the context in which they occur. Findings were first presented and then analyzed in terms of the amount to which the searched items contributed to transmitting specific persuasive and/or manipulative ideas following the three stages of analysis.

C. The Frequency Distribution Analysis

The work of concordance was limited to the analytical process, by providing an FDA for the searched lexical elements that were identified as significant in the study of modality as ideology indicators. Concordance is used to conduct this frequency analysis. Concordance makes it easier to access and examine big data texts in order to produce reliable and succinct results that would be impossible to achieve if the analysis were done without the aid of computer tools [37-39]. In this case, the concordance options simply allowed you to mark the word in its context. This serves as a quick summary of the linguistic context in which the term appeared in the text. Concordance, according to Kennedy [20], is software that generates all instances of a given word or lexis in a corpus. Furthermore, according to Hockey [19], a concordance or frequency analysis is formed by the searched item and the context in which it occurs. For example, Concordance provides KWIC (Key Word in Context), which provides a wealth of information about the searched word in its many contexts in text. As a result, the interpretative process is extended, opening up new perspectives that aid in better understanding the language representation.

IV. Data Analysis and Results

This section presents an FDA of the modal verbs under investigation. The modal verbs will analytically be divided according to their semantic functionality into three categorizations: first, the modal verbs that indicate truthfulness and certitude (will); and, second, the modal verbs that communicate obligation (must). This will be accompanied by a content analysis and a key word in context analysis to clarify the ideologies carried by these modal verbs in the play under investigation.

A. An FDA of the Truth Modals

The truth modal ‘will’ is employed by many characters in the discourse of the play to emphasize their credibility and

demonstrate the veracity of their words The modal 'will' is utilized in Lear to convey both persuasive and manipulative ideologies. In more than one situation in the discourse of Lear, both Fontanelle and Bodice tried to convince their father, the old king, to permit and approve their furtive marriage from his ancestral foes, the duke of North and the duke of Cornwall, as mentioned in the preceding extract. They both know their father would reject their marriage, so they utilize the truth modal 'will' to sway their father's attitude toward their husbands. Fontanelle's statement, "I know you'll get along with my husband," expresses her desire to marry Cornwall. Bodice's remark that she'll soon learn to respect them as if they were her sons is yet another attempt to persuade her father to approve her marriage from the North. Bodice's usage of the pronoun 'you' demonstrates her authority in communicating with her father. The two daughters attempt to eliminate Lear's fear of the two husbands from his thoughts so that he will accept their marriage without reservation. A frequency distribution analysis of the manipulative and persuasive 'will' is presented in Tables I and II.

TABLE I. A FREQUENCY DISTRIBUTION ANALYSIS OF MANIPULATIVE 'WILL'

The Modal Verb	Total Frequency	Indicative Occurrences
Will	77	5

TABLE II. A FREQUENCY DISTRIBUTION ANALYSIS OF PERSUASIVE 'WILL'

The Modal Verb	Total Frequency	Indicative Occurrences
will	77	4

Tables I and II demonstrate that the truth modal 'will' has a total frequency of 77 occurrences; only 9 occurrences are significant in conveying specific meanings and ideologies, 5 of which are utilized to convey manipulative ideology (Table I), and 4 occurrences are employed to channel persuasive ideology (Table II). The complementary link between the two CATA variables employed here, the FDA and the KWIC variables, is further highlighted in these two tables. To further elucidate this point, it is clear that, despite its ability to provide us with the total frequency of a certain word, FDA is still unable to assist us in better understanding the indicative occurrence of that term. Only by using the KWIC variable can one determine what is suggestive and what is not among occurrences. The two factors' complimentary nature discursively supports the entire interpretative atmosphere of the play under investigation.

B. An FDA of the Obligation Modals

It is grammatically known that the modal 'must' is recurrently used to communicate the grammatical function of obligation. This obligation modal is deftly deployed in Bond's Lear to depict the speaker's authority over his audience. Obligation modals are used by speakers to impose their own ideology on their listeners and to drive their conduct toward perfect compliance and subordination to their goals [40]. The employment of the obligation modality dominates oppression discourse, in which powerful characters use these modals to exercise dominance over the powerless.

Bodice and her sister, Fontanelle, are discussing their plot to attack Lear's army and evict him from his throne. Bodice use the obligation modal 'must' three times to highlight the importance of accomplishing what they plan to do, as well as the urgency of attacking their father in order to stop the acts of building on the wall. We must travel to our spouse, we must attack before the wall is done, and we must aid each other are all obligation modals used by Bodice to highlight her authority and dominance over her sister. Even in her relationship with her spouse, she directs her. The pronoun 'we,' which accompanies the modals, reflects unity, which Bodice tries to convey to Fontanelle in order to ensure that she is looking out for her sister's best interests; this, in turn, pushes Fontanelle to willingly carry out what her sister demands. As a result, the obligation modal 'must' is perverted in order to channel deceitful ideology. The frequency analysis that follows provides greater insight into the manipulative use of the obligation modal 'must' in both affirmative and negative variants.

TABLE III. A FREQUENCY DISTRIBUTION ANALYSIS OF 'MUST'

The Modal Verb	Total Frequency	Indicative Occurrences
must	100	13

TABLE IV. A FREQUENCY DISTRIBUTION ANALYSIS OF 'MUSTN'T'

The Modal Verb	Total Frequency	Indicative Occurrences
Mustn't	9	1

Tables III and IV show that in the novel's discourse, 13 occurrences of the affirmative 'must' and 1 occurrence of the negative 'mustn't' are used as bearers of manipulative ideology. The negative obligation modal, although its rarity, is very indicative in transmitting manipulative views. The fact that a term appears frequently does not necessarily mean that it is thematically suggestive. Low frequency terms, on the other hand, are often quite suggestive.

V. DISCUSSION

The analysis demonstrates that the modal verbs of truth (will) and obligation (must, mustn't) Go beyond their grammatical and semantic duties to express and sustain certain pragmatic and ideological meanings, such as persuasion and manipulation, at times and in specific settings. In the discourse of Bond's Lear, the modal verbs under consideration express a distinct form of ideology. Truth and obligation modalities are particularly important in expressing both persuasive and manipulative ideas in the play's discourse, according to analytical data.

The analysis demonstrated that the use of CATA software aids in the extraction of new meanings and ideologies that are supposed to be unclear to the ordinary reader. This computational approach to textual analysis further serves to achieve better understanding of the significant role of modal verbs as conduits of particular meanings. It also functions to enhance the discursively analytical amalgamation of discourse studies and computers, especially in the linguistic analysis of massive data texts. The two variables FDA and KWIC are harmonizing in nature, because the latter is a context-oriented

variable that focuses on the recognition of significant words created by the former's total occurrences realized by virtue of an FDA. Both FDA and KWIC provide significant contributions to literary text linguistic analysis, notably in deciphering latent ideologies beyond the meaning propositions of plain language phrases.

The study establishes that modal verbs in language have ideological meaning. This is consistent with Fowler's [29] notion that language and ideology have a reciprocal relationship, because each linguistic expression (i.e., word, phrase, and sentence) can express the user's individual ideology. Ideology is usually present in language, and the use of specific verbal expressions over others has an ideological basis within textual analysis. That is, it is generated in this specific fashion and with this unique linguistic expression to communicate the speaker/specific writer's ideological connotations. As a result, every single word might reflect the user's philosophy. It is not simply content words that express and preserve ideas in conversation in the context of this study.

The analysis further clarified that function words, on the other hand, play an important role in communicating and maintaining ideals. Function words lose their common semantic meaning in certain discourse circumstances in order to express additional ideological goals. Modality, either classified as obligation or categorized as truth, is also used to portray deceptive ideology. The obligation and truth modal verbs are also employed to convey necessity and certainty. In the oppressive discourse, all of these methods are more representative. This is consistent with prior research [5], [11], and [41], which emphasize the ideological significance of modality.

The current linguistic investigation also reveals that modality in Bond's *Lear* is used to deliver persuasive ideas (must, will). The modals obligation and truth are used to convey requirement and certainty. These modal verbs are used to present a convincing ideology that is founded on facts and previous experiences. This aligns with the argument of Sornig [42], who postulates that within ideological discourse, the persuasive purposes beyond discourse can be deciphered by a variety of linguistic tools, including semantic, pragmatic and grammatical devices. This, for him, emphasizes the fact that the different levels of linguistic analysis can analytically be integrated towards the realization of the intended meaning of the speaker/writer. Thus, the linguistic investigation of texts can be conducted at the syntactic, semantic, pragmatic and/or grammatical levels.

VI. Conclusion

The current study used a computer-assisted text analysis to decipher the ideological importance of modal verbs (will, must, and mustn't) in Edward Bond's Lear dialogue. The study employed two analytical approaches: first, critical discourse analysis, as described by Fairclough's model of analyzing discourse in terms of its grammatical features; and second, computer-aided text analysis, which is analytically enabled by CATA's three variables: frequency distribution analysis (FDA), key word in context (KWIC), and content analysis (CA). The three techniques are analyzed to see how much each modality contributes to the communication of specific beliefs in the selected text, ranging from persuasion to manipulation. The use of modality for ideological reasons was demonstrated in the analysis of the chosen play. The analysis also clarified that the various ideological meanings in discourse can be exposed by the skillful usage of modal verbs. These modal verbs in particular contexts cease to convey their ordinary semantic purpose of, for example, agency, certitude, obligation, etc., and channel further and new meanings either at the character-to-character or author-to-reader levels of communication. The analysis further revealed that modal verbs in light of this article are utilized used to bring about two types of ideology: manipulative and persuasive; the former always serves the interests of the speakers and/or writers, whilst the latter frequently functions to address the interests of all participants involved in any communicative acts.

Finally, this study suggests that other CATA variables, such as LWIC (Linguistic Inquiry and Word Count) and DICTION (software package with 31 predefined Dictionaries), be applied to the textual, pragmatic, stylistic and thematic analysis of other types of function words, such as the ideological analysis of prepositions, conjunctions, and demonstratives. This could lead to different and/or comparable conclusions than those presented in this research. The report also suggests that CATA software be used in the field of EFL teaching and learning, notably in literature courses. This may result in improved student learning outcomes as well as the development of innovative teaching approaches on the side of teachers.

Acknowledgment

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

References

[1] J. Reddington, F. Murtagh, and C. Douglas, "Computational properties of fiction writing and collaborative work," International Symposium on Intelligent Data Analysis, pp. 1-13, 2013.

[2] A. F. Khafaga, and I. Shaalan, "Using concordance to decode the ideological weight of lexis in learning narrative literature: A computational approach," International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, pp. 246-252, 2020.

[3] K. Beatty, Teaching and researching computer-assisted language learning. Harlow: Longman Pearson, 2010.

[4] A. Thabet, "Applied computational linguistics: An approach to analysis and evaluation of EFL materials," Damietta Faculty of Education Journal, vol. 1, no. 13, pp. 7-39, 1990.

[5] A. F. Khafaga, and I. Shaalan, "Pronouns and modality as ideology carriers in George Orwell"s Animal Farm: A computer-aided critical discourse analysis," TESOL International Journal, vol. 16, no. 4.2, pp. 78-102, 2021.

[6] M. Eltahir, S. Al-Qatawneh, and S. Alsalhi, "E-Textbooks and their application levels, from the perspective of faculty members at Ajman University, U.A.E," International Journal of Emerging Technologies in Learning, vol. 14, no. 13, pp. 88-104, 2019.

[7] G. Stockwell, Computer-assisted language learning: Diversity in research and practice. Cambridge: Cambridge University Press, 2018.

[8] J. Sinclair, Corpus, concordance collocation. Oxford: Oxford University Press, 1991.

[9] A. F. Khafaga, Strategies of political persuasion in literary genres: A computational approach to critical discourse analysis. Germany: LAMBERT Publication, 2017.

[10] D. Wiechmann, and S. Fuhs, "Concordancing software," Corpus Linguistics and Linguistic Theory, vol. 2, no. 2, pp. 107-127, 2006.

[11] N. Fairclough, Language and power. London and New York: Longman, 1989.

[12] E. Bond, Lear, In plays two. London: Eyre Methuen, 1978.

[13] C. Pim, "Emerging technologies, emerging minds: Digital innovations within the primary sector," in G. Motteram (Ed.), Innovations in learning technologies for English language teaching, London: British Council, 2013, pp. 17-42.

[14] J. Jarvis, and L. Pastuszka, "Electronic literacy reading skills and the challenges for English for academic purposes," CALL-EJ Online, vol. 10, no. 1, 2008.

[15] S. Pinner, "Teachers' attitudes to and motivations for using CALL in and around the language classroom," Procedia-Social and Behavioral Sciences, 34, pp. 188-192, 2012.

[16] R. Fowler, and G. Kress, "Critical linguistics," in Fowler, R., Hodge, R., Kress, G., and T. Trew, T., (Eds). Language and control. London: Routledge and Kegan Paul, 1979, pp. 185-213.

[17] T. A. van Dijk, "Ideological discourse analysis," in Ventola, E., and Solin, A. (Eds.), Interdisciplinary approaches to discourse analysis. New Courant, pp. 1995, 135-116.

[18] Q. Ma, "From monitoring users to controlling user actions: A new perspective on the user-centred approach to CALL," Computer Assisted Language Learning, vol. 20, no. 4, pp. 297-321, 2007.

[19] S. Hockey, A guide to computer applications in the humanities. London: The Johns Hopkins University Press, 1980.

[20] G. Kennedy, An introduction to corpus linguistics. London & New York: Longman, 1998.

[21] A. Barger, and K. Byrd, "Motivation and computer-based instructional design," Journal of Cross-Disciplinary Perspectives in Education, vol. 4, no. 1, pp. 1-9, 2011.

[22] Y. H. Chen, "Computer mediated communication: The use of CMC to develop EFL learners' communicative competence," Asian EFL Journal, vol. 7, no. 1, pp. 167-182, 2005.

[23] G. Stockwell, "Computer-assisted language learning: Diversity in research and practice. Cambridge: Cambridge University Press, 2018.

[24] K. Romeo, "A web-based listening methodology for studying relative clause acquisition," Computer Assisted Language Learning, vol. 21, no. 1, pp. 51-66, 2008.

[25] R. Dzekoe, "Computer-based multimodal composing activities, self-revision, and L2 acquisition through writing," Language Learning & Technology, vol. 21, no. 2, pp.73-95, 2017.

[26] J. Flowerdew, "Concordancing as a tool in course design," System, vol. 21, no. 2, pp. 231-244, 1993.

[27] D. Krieger, "Corpus linguistics: What it is and how it can be applied to teaching," The Internet TESL Journal, vol. IX, no. 3, pp. 123-141, 2003.

[28] T. A. van Dijk, "Discourse, knowledge and ideology: Reformulating old questions and proposing some new solutions, in Martin, P., Aertselaer, J. N. and van Dijk, T. A., (Eds.), Communicating ideologies: Multidisciplinary perspectives on language, discourse, and social practice. New York & Oxford: Peter Lang, 2004, pp. 5-38.

[29] R. Fowler, "On critical linguistics," in Coulthard, C., and Coulthard, M., (Eds.), Texts and practices: Readings in critical discourse analysis**.** London & New York: Routledge, 1996, pp. 15-31.

[30] T. A. van Dijk, "Ideology and discourse analysis," Journal of Political Ideologies, vol. 11, no. 2, pp. 115-140, 2006.

[31] R. Fowler, Language in the news: Discourse and ideology in the press. London: Routledge, 1991.

[32] A. F. Khafaga, "Exploring ideologies of function words in George Orwell's Animal Farm," Pertanika Journal of Social Sciences and Humanities, vol. 29, no. 3, pp. 2089 -211, 2021.

[33] F. Yavus, "The use of concordancing programs in ELT," Procedia-Social and Behavioral Sciences, 116, pp. 2312-2315, 2014.

[34] A. Omar, A. F. Khafaga, and I. Shaalan, "The impact of translation software on improving the performance of translation majors," International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, pp. 287-292, 2020.

[35] A. F. Khafaga, and A. Alghawli, "The impact of CALL software on the performance of EFL students on the Saudi university context," International Journal of Advanced Computer Science and Application, vol. 12, no. 7, pp. 304-312, 2021.

[36] A. F. Khafaga, and B. Aldossari, "The language of persuasion in courtroom discourse: A computer-aided text analysis," International Journal of Advanced Computer Science and Application, vol. 11, no. 7, pp. 332-340, 2021.

[37] M. L. Heyden, J. Oehmichen, S. Nichting, and H. W. Volberda, H., "Board background heterogene-ity and exploration-exploitation: The role of the institutionally adopted board model," Global Strategy Journal, vol. 5, no. 2, pp. 154–176, 2015.

[38] A. F. Khafaga, "A computational approach to explore the extremist ideologies of Daesh discourse," International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 193-199, 2020.

[39] W. Abraham, Modality in syntax, semantics and pragmatics. Cambridge: Cambridge University Press, 2020.

[40] A. Khafaga, "Linguistic manipulation of political myth in Margaret Atwood's The Handmaid's Tale," International Journal of English Linguistics, vol. 7, no. 3, 189-200, 2017.

[41] A. F. Khafaga, "Linguistic representation of power in Edward Bond's Lear: A lexico-pragmatic approach to critical discourse analysis," International Journal of English Linguistics, vol. 9, no. 6, 404-420, 2019.

[42] K. Sornig, "Some remarks on linguistic strategies of persuasion," in Wodak, R. (Ed.), Language, power and ideology: Studies in political discourse. John Benjamins Publishing Company,1989, pp. 95-113.

Detection of Criminal Behavior at the Residential Unit based on Deep Convolutional Neural Network

H.A. Razak[1], N.K. Zakaria[4], N.F.M. Zamri[5]
College of Engineering, Universiti Teknologi MARA
Selangor, Malaysia

Nooritawati Md Tahir[2]*
College of Engineering
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia
Institute for Big Data Analytics and
Artificial Intelligence (IBDAAI),
Universiti Teknologi MARA
Selangor, Malaysia

Ali Abd Almisreb[3]
Faculty of Engineering and Natural Sciences
International University of Sarajevo
Sarajevo, Bosnia-Herzegovina

***Abstract*—Studies on abnormal behavior based on deep learning as a processing platform increase. Deep learning, specifically the convolutional neural network (CNN), is known for learning the features directly from the raw image. In return, CNN requires a high-performance hardware platform to accommodate its computational cost like AlexNet and VGG-16 with 62 million and 138 million parameters, respectively. Hence in this study, four CNN samplings with different architectures in detecting abnormal behavior at the gate of residential units are evaluated and validated. The forensic postures, with some other collected data, are used for the preliminary step in constructing the criminal case database. High accuracy up to 97% is obtained from the trained CNN samplings with 80% to 97% recognition rate achieved during the offline testing and 70% to 90% recognition rate recorded during the real-time testing. Results showed that the developed CNN samplings owned good performance and can be utilized in detecting and recognizing the normal and abnormal behavior at the gate of residential units.**

***Keywords*—Abnormal behavior; deep learning; convolution neural network; forensic posture; property crime**

I. Introduction

There is an increase in the usage of closed-circuit television (CCTV) in residential units as a consequence of the upbringing awareness of sheltered zone [1]–[3]. This monotonous observation can cause fatigue and distraction, leading to negligence and being overlooked as the surveillance process is underway [4]. Currently, numerous studies are conducted to detect and track objects and people's anomalous state in developing intelligent surveillance systems [4]–[7], which is related to the changing pattern or movement of objects or humans from the original form or behavior, referred as anomalous. At present, image recognition is the most appropriate technology to utilize CCTV footage optimally, and the recognition can yield better results using deep learning techniques.

Deep learning is a hierarchical feature learning to classify multidimensional and complex data set elements. There are several types of deep learning structures that include convolutional neural network (CNN), long-short term memory (LSTM) and recurrent neural network (RNN). CNN is suitable for object detection and image recognition and has been widely used in numerous biometrics applications, namely fingerprint, iris, and gait [8]–[10]. Recall that gait biometrics can be used to identify subjects from their style or manner of walking. Due to its uniqueness, gait is also considered competent biometrics, suitable for forensic intelligent surveillance systems. This is because gait as biometric has the potential for farther distance recognition, can be possessed without the perpetrators' consent and awareness, and can also be perceived at a low-resolution camera. Combining these technologies, image recognition, CNN, and gait biometric brings us a little closer to developing a forensic intelligent surveillance system. However, the lack of data on criminal behavior in public databases leads to the problem of developing and designing adaptability features of forensic gait for recognition and detection.

Hence, the main objective of this study is to investigate and validate the forensic postures with the authorities' consent in interpreting anomalous behavior during the housebreaking crime in residential units. Four CNNs with different architectures, namely Up, Down, Up-Down, and Down-Up sampling, are developed to classify massive data on both normal and anomalous behavior. The effectiveness of the developed CNN samplings is examined with two test modes, offline and real-time detection. The offline detection is evaluated using several CCTV footage of the actual housebreaking crimes, and the real-time detection is held at the laboratory. It is essential to understand the architectures of CNN to develop a robust network at a minimum computational cost that can be a kick start in developing a robust and economical forensic intelligent surveillance system.

*Corresponding Author.
This research is funded by the Ministry of Higher Education (MOHE), Malaysia via the Fundamental Research Grant Scheme (FRGS) No: FRGS/1/2019/TK04/UITM/01/3.

II. Related Work

As mentioned earlier, CNN can be considered for gait recognition due to its outstanding realization. Currently, CNN comes in handy whenever needed to classify the image dataset without going through the pre-processing image and feature extraction step. These steps are handled by the multiple layers of nonlinear where the output from previous layers is the input to the following layers. CCN is able to learn automatically the significant features of large input image databases based on the pre-defined size of each filter in the convolution layers that establishes the convolution map uniformly [11]. The pooling layers minimize the redundant pixels based on the rescaling map, further reducing the feature matrix size of the convolution maps [12]. As for the convolution layer, the stochastic gradient descents with momentum (SGDM) optimization function and the activation function of the rectified linear unit (ReLU) are used accordingly during the learning process. This process continues with the features subset connected with each other that further developed the connection of the classification output layer. This is achieved as the first convolution maps preserved feature vectors provides learning to the second convolution layer [13]. For CNN, note that numbers of parameters are decreased accordingly during the convolution and pooling process, specifically the connections and shared weights [14].

As reported in [15], CNN is used for tracking humans based on numerous poses, viewpoints including occlusion, using ten challenging datasets. Here, CNN with five layers were developed that consists of convolutional, pooling, normalization, fully-connected and softmax layer. The 4 by 4 by k channels with two strides and zero-padding and 50 filter banks were used as the convolutional layer. As for the pooling layer, the filter size is 2 by 2 with a max operator and two strides and zero padding. For the normalization layer, the pre-defined hyperparameters are set as k=1, ρ=2, α=1/4, and β=0.5. Next, the fully-connected layer flattened the extracted features associated with the softmax node. The classification performance was evaluated using the softmax operator and the log loss values with fixed hyperparameters during training; five as the maximum epoch, 0.001 as the learning rate, and ten as the batch size. The developed CNN was evaluated to track variation of occlusion using the women dataset since this dataset comprised numerous poses with partial occlusion of the lower and upper limb. Next, the basketball game video was used for testing body deformation variation. By averaging the Euclidean distance of the frames ground-truth positions and the person being tracked, the error of the centre location acted as the tracking result. Result attained was 91.31% using the basketball dataset and for the women dataset was 94.14%.

Conversely, as reported in [5], malicious activities were investigated for three anomalous behaviors based on six CNN layers, three convolutional layers, two fully connected layers, and one softmax layer. Filter size and total filters were fixed for all three convolutional layers. The same goes for the convolution stride, pooling and ReLU. The pooling layers utilized the max operator. Two experiments were conducted. The first fully-connected layers have 64 neurons as output, with the output neurons set as two and six accordingly for each experiment for the second fully-connected layer. Further, between the two fully-connected layers a ReLU layer was added. For each category, the probability was computed in the softmax loss layer. SGDM was used as the network optimization function with the learning rate range set from 10^{-3} to 10^{-1} whilst the epoch was set from 10 up to 100. Further, the developed CNN was tested using images based on five datasets comprised of normal and anomalous behaviors and variations. Ratio for training and testing was set as 70:30. However, for the PEL dataset, since most of the dataset consists of fighting scenes acquired from movies, the dataset was split as 43:57 as training and testing. Firstly, the algorithms classified the datasets into two categories; normal and abnormal, followed by classifying the abnormal category into three different anomalous behaviors. The three anomalous behaviors are punching, pushing and kicking. For all datasets, it was found that higher accuracy is achieved as the epoch is increase although more time is required for the learning rate to finally stabilize at 0.001. The accuracy attained was 100% for anomalous behaviors detection for both experiments based on the developed CNN.

Moreover, research on human behavior has gained the attention for safety community purposes, especially the large-scale industry since they are dealing with dozens and even hundreds of employees and equipment every day. As reported in [16], temporal information of human activities in the industry of each frame was processed using motion history image (MHI) and discrete cosine transform (DCT) to generate the 2D spatial-temporal maps. This process has successfully reduced the size of each frame from 704×576 to 88×72 with minimal information loss. These 2D maps were fed to the CNN for identifying human behavior and activity in the industrial environment. The developed CNN consists of three convolutional layers and one output layer of multi-layer perceptron (MLP). The first two convolutional layers have similar size with regards to the filter and zero padding. The number of filters in the second convolutional layer was 40, and it doubled from the first layer, which were 20 filters. The third convolutional layer had 60 filters with the dimension of 5×5. A pooling layer followed each convolution layer. Next, the final pooling layer flattened the convolutional feature vectors forming 1440-dimensional feature vectors. These feature vectors were inputs of 600 neurons of the first hidden layer of MLP and the output layer of six neurons representing the number of behaviors or activities to be classified. SCOVIS dataset contains heavy occlusion, the interaction between humans and machinery and factory environments were suitable to validate the network. 15 scenarios for training and 5 scenarios for testing were used for the SCOVIS dataset. Precision and recall were almost 99% for the training set and nearly 90% for the unseen dataset.

Recently, studies on anomalous behavior during driving environments have been very encouraging. As discussed in [17], deep learning based on CNN architectural was developed, known as DedistractedNet, that was used to classify the distracted driving behaviors like texting, drinking, putting on cosmetics and many more. This network has five sets of a convolutional layer, ReLU and max pooling layer, followed by fully connected layers with neurons corresponding to eight driver behaviors. The cross-entropy loss function computed the

category loss of DedistractedNet. The learning process acquired 9840 images, with 9120 images used for training and 720 images for testing. The network was compared with two pre-trained CNN, LeNet and AlexNet. The results of similarity and F1-measure showed that DedistractedNet preceded both LeNet and AlexNet in all categories.

Based on previous work and findings by [14] that detected the criminal behavior using CNN as motivation in this work, we aim to investigate the forensic postures on anomalous human behavior at the gate of residential units as the database. Next is to develop four samplings of CNN with different architectures operated on a humble hardware platform.

III. Architecture of Convolution Neural Network

Hubel and Wiesel found that the neuron cells in the visual cortex of the cats were able to produce visual perceptions through self-organizing the image structure by learning from the experiences [18]. These cells are sensitive to a specific visual field region that is generally referred to as perceptive fields [19]. The specific tasks of the neuron cells in the visual cortex have been considered the genesis behind CNN's invention. Applying the same idea, let $F \in {}^{m_H \times n_W}$ and $W^k \in \mathbb{R}^{m \times n_{l-1}}$ be a matrix representing the perceptive field and neuron cell or filter size and weights in machine learning, respectively. In this study, the input type for CNN is an order 3 tensor, $X \in \mathbb{R}^{n_H \times n_W \times n_C}$ represents an image with *H* rows of neuron, *W* columns of neuron and *C* of color channels.

A. Convolution Layer

The convolution layers are characterized by an input map, *X*, a bank of independent filters or kernel, *F*, and biases, *b*. Each filter is convolved individually with the input map to produce a feature map, ϕ. indicated the relationship between the input and output of the network. *Y* can be written as, $Y = \varphi(WX + b)$ where *X* is the inputs, *Y* is the output, *W* is the weights, *b* is the biases, and φ is the activation parameters.

The feature map, ϕ is defined as the relationship between the input image $X \in \mathbb{R}^{n_H \times n_W \times n_C}$ and the filter or kernel, $F \in \mathbb{R}^{n_H \times n_W}$.

$$\phi = X \otimes F, \ \forall W \in F \tag{1}$$

Let the convolution of input feature maps, $X \in \mathbb{R}^{n_H \times n_W \times n_C}$ with a bank of *D* multi-dimensional filters, $F \in \mathbb{R}^{n_H \times n_W \times n_C \times D}$ and biases, $b \in \mathbb{R}^D$ one for each filter. The output from the convolving process of the input *X* by transposing the filter to implement data interpolation is given by the convolution theorem [20][21].

$$\left(x \cdot w\right)_{ij} = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} \sum_{c=1}^{C} w_{m,n,c} \cdot x_{i+m,\, j+n,c} + b \tag{2}$$

In essence, the convolution layer is a hierarchical model that comprises multiple convolution layers to train massive data in achieving the highest accuracy in detection and recognition. As presented in Fig. 2, there are two stages in the learning procedure of the convolution layer namely forward and backward propagation or also known as back propagation.

Forward propagation procedure calculates the output *z*, using the inputs, *x*. Meanwhile, the back propagation procedure takes the gradient of the loss function concerning output, ∇L_z as the input of the network and the gradients of *x* for the loss function, ∇L_x to implement the updating procedure of weights through convolution layers. The back propagation algorithm utilizes the effectiveness of the chain rule in handling the derivatives recursively to obtain the desired weights. Understanding that the network's input is a single vector, and the convolution layer consists of a set of neurons, each vector will convolve with each neuron during the learning process on the convolution layer. It is appropriate to understand the dimensional vector in each layer of the network in building the CNN.

There are two weaknesses during the convolution procedure. Firstly, it shrinks the image, and secondly, it discards great numbers of information near the edge of the image. Hyperparameters are introduced to solve these problems, first is *f* as the filter size that generally in odd size to attain symmetrical padding. Next is *p*, the padding by adding columns and rows of zero to preserve the spatial sizes of feature maps, and finally parameter *s*, known as the stride, the number of pixels that move when traversing the input during convolution. The convolution layer has various feature map sizes as the hyper parameters can modify it. Generally, feature maps and output volume can be expressed as, $\phi(z) \in \mathbb{R}^{n_H \times n_W}$ and $z \in \mathbb{R}^{n_H \times n_W \times n_C}$ respectively.

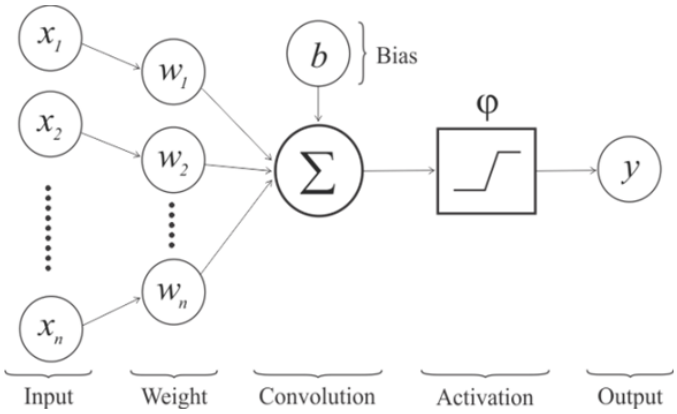


Fig. 1. The Architecture of Neural Network.

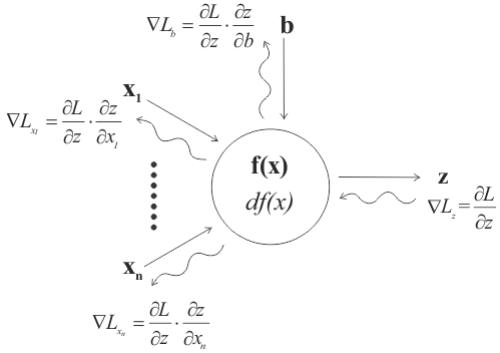


Fig. 2. Change in Vector during One Iteration of forward and backward Propagation.

B. Optimization Function

Calculus helps the learning process in machine learning to improve prediction accuracy by calculating the derivatives during the optimization procedure. The most common optimization function in CNN is gradient descent. The hallmark of gradient descent is required only the first order of derivatives of parameters concerning the loss function. The lower error value in the loss function demonstrated that better predictions had been calculated for the network. The Stochastic Gradient Descent (SGD) trains the learning algorithm to minimize errors, the calculation of the slope of error towards the negative of the gradient to find the global minima of the network. However, the downside of SGD is it complicates the convergence to potentially better global minima because the error rates keep overshooting due to the frequent updates. It results in SGD being computationally expensive and highly ineffective for memory, making SGDM a popular choice in the learning process of CNN. The velocity and friction parameters, β, applied in SGDM can prevent overshooting while allowing faster convergence. Adding the SGDM parameters into the gradient of the loss function to weight ∇L_w allows updating the consequences in the network. Furthermore, the parameters can steer the gradient vectors to accelerate in the right direction with the knowledge of the previous surface curve in the ravine [22]–[24]. In this study, the friction, β, is set to 0.9.

C. Activation Function

The primary purpose of the activation layer is to convert the input map of each neuron to the output feature map, which will then be used as the input map in the next layer. Essentially, the non-linear activation function is differentiable. Therefore, it allows the back propagation optimization technique to reduce errors by optimizing the weights using gradient descent [25]. Additionally, the function enables the neurons to learn the complex functional mapping from input data due to its curvature quality, considering the function has more than one degree. Traditionally, the sigmoid and hyperbolic tangent has been broadly used for the neural network but become irrelevant for many layers networks due to the vanishing gradient problem and slow convergence [25], [26]. The most appropriate activation function for CNN is rectified linear unit (ReLU) [27]. The advantages of ReLU are sparsely activated, that offers better predictive power and lessen over fitting to the training set, faster converging, avoids vanishing gradient problem, and the best attribute is computational economical as it excludes complicated mathematic functions. The ReLU mathematical expression can be written as [25][28];

Let $\varphi_k : \mathbb{R} \to \mathbb{R}$ be a non-linear activation function.

$$\varphi_k = max\begin{Bmatrix} 0 & for & x < 0 \\ x & for & x \geq 0 \end{Bmatrix}^{\infty}_{0} \tag{3}$$

D. Pooling Layer

The pooling layer role reduces the spatial dimension of input volume for the next layer. Therefore, the hyper parameter of the pooling layer is the stride only. Padding is rarely applied in this layer. The depths of this layer also remain the same, $n_C^{[\ell]} = n_C^{[\ell-1]}$. The feature map of the output can be calculated as follows,

$$\phi(z) = \left(\frac{n_H^{[\ell-1]} - f_H^{[\ell]}}{s}\right) \times \left(\frac{n_W^{[\ell-1]} - f_W^{[\ell]}}{s}\right) \tag{4}$$

E. Fully-Connected Layer

The fully-connected layer requires a vector as the input, therefore the flattening procedure is performed towards n-dimensional vectors where n >1. The flattening procedure is transforming spatial structure data into one dimensional feature vectors by concatenating the tridimensional tensor of convolution layer output into the monodimensional tensor that is a vector. The vectors are learned using gradient descent to ensure the class scores are accordant with the labels in the learning set of each image.

F. Classification Layer

The properties of the classification layer are the softmax activation function and cross-entropy loss function. The softmax activation function is generally applied to the final layer of the networks. It calculates a probabilistic value for every class between 0 and 1. The cross-entropy loss function measures the optimization for multi-class classification.

The softmax activation function has excellent features. Firstly the normalized data increases consistency and convenience in mapping. Next, it is differentiable and practicable in calculating the loss function. Finally, it employs the natural exponential that owns the ability to identify the difference between higher and lower values. In CNN, the softmax activation function takes the actual values of a K-dimensional vector on both input and output vectors, transforms it to a range of between 0 and 1, which is essential for the probability distribution. The softmax activation function at the output layer of the network can be defined as;

$$\sigma_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}, \quad \forall j \in 1 \ldots K \tag{5}$$

Hence, the vector will be trained using cross-entropy loss function, ζ, to predict which input vector belongs to one of the output classes.

$$\zeta(x,y)_j = \sum_{j=1}^{K} x_j \ \ell n \, y_j \tag{6}$$

IV. Proposed CNN Samplings

Recall that this study proposes to detect anomalous behavior at the gate of residential units using from-scratch CNNs on a standard notebook as the hardware platform. The idea of having an extensive network was to gain a more robust network with large training datasets to satisfy the nonlinear algorithms requirement. This improves the network's skill and avoids overfitting, making the CNN computationally expensive to solely operate on GPU. Four types of CNN sampling with nearly twenty thousand images as training datasets have been utilized to defy the odds.

The CNN module named sampling refers to the kernel's change in size and depth or filter hyperparameter. However, the other hyperparameters of the convolution layer are set to a fixed value. The four modules are Up sampling, Down sampling, Up-Down sampling, and Down-Up sampling. There are three sizes of the filter being employed for the experiments as in Table I.

TABLE I. Value of Hyperparameters for Convolution Layers

Layers	**Padding, *p***	**Stride, *s***	**Filter Size, *f***	**No. of filter, n_c**
Conv1	1	1	3 x 3	16
Conv1	1	1	7 x 7	32
Conv1	1	1	11 x 11	64

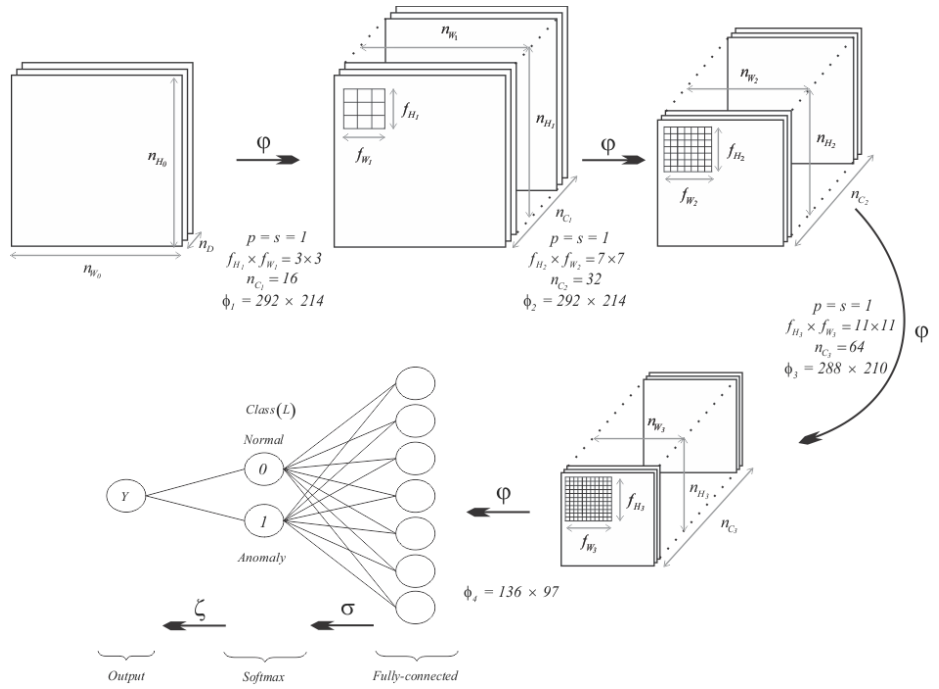
It is necessary to set the variables or hyperparameters of the training algorithm before executing each module as listed in Table II. Towards comparing each module, all hyperparameters values of the training procedure have been set to the same value. All sampling types consist of a convolution layer that alternates with the ReLU layer, the pooling layer and the networks end with the fully-connected layer followed by the softmax layer. The Up sampling is arranged in the ascending order of convolution layers namely Conv1, Conv2 and Conv3. Meanwhile, the Down sampling is organized in descending order specifically Conv3, Conv2, Conv1. As for Up-Down sampling, it starts with an ascending order and followed by descending order of the convolution layer that are Conv1, Conv2, Conv3, Conv3, Conv2, and Conv1 and the Down-Up sampling is in reverse designed with descending order at the beginning and continued by ascending order of convolution layer namely Conv3, Conv2, Conv1, Conv1, Conv2 and Conv3. In this study, the hyperparameter padding and stride of the convolution layer are set to 1 as in Table I. However, the hyperparameter stride of pooling layer can be varied. The architecture of CNN for each sampling is shown in Fig. 3. Next, Table III presents the detailed network configuration information on each developed sampling type in this study.

TABLE II. Hyperparameter for CNN Samplings

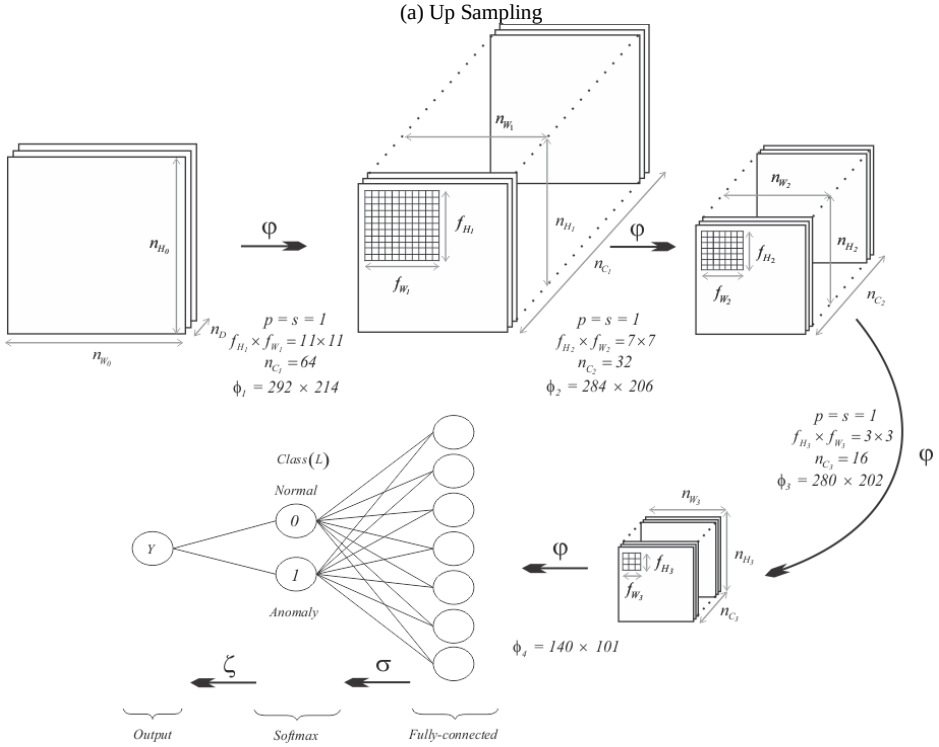
Hyper-parameter	**Size**	**Task**
Learning type	SGDM 0.9	• To prevent oscillations • To navigate the gradients towards optimum global minima
Learning rate	0.001	• To determine the speed of updating the trainable parameters • To assist in constant converging
Epoch	1000	• To start network propagates in both forward and backward • To activate the neuron • To calculate the loss • To obtain the partial derivative of the loss function • To update trainable parameters
Minibatch size	20	• To achieve training stability • To improve performance
Validation frequency	50	• To validate the network at regular interval

TABLE III. Network Configuration of all Samplings

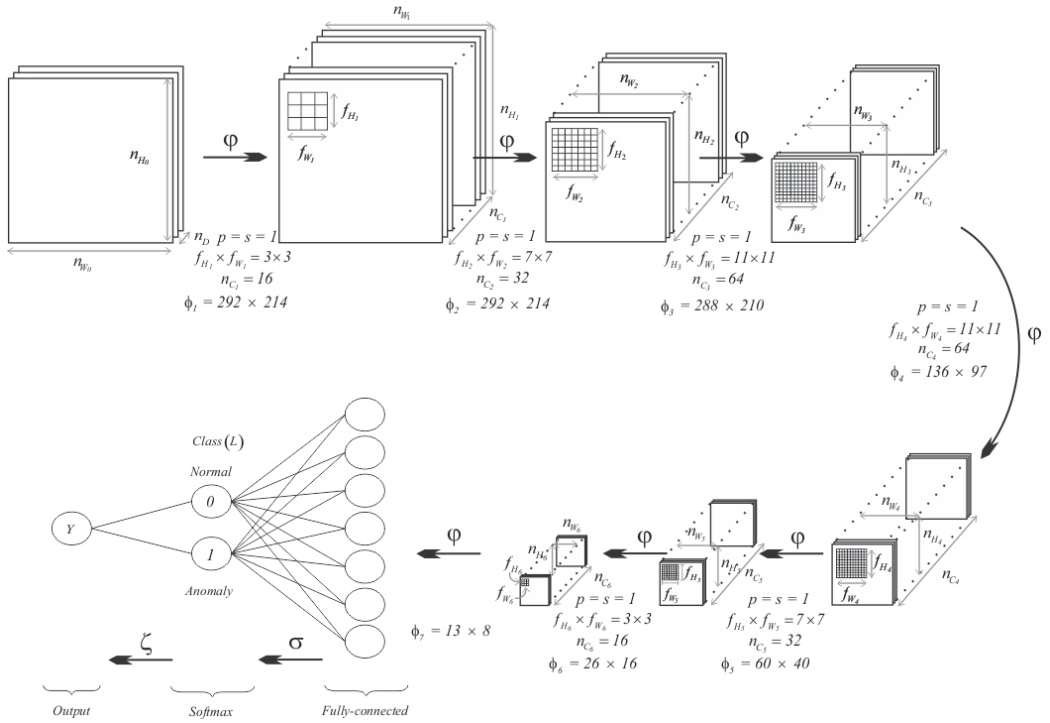
Layers	**Up**	**Down**	**Up-Down**	**Down-Up**
Total layers	15	15	27	27
Image size	292 x 214 x 3			
Convolution 1	3 x 3 x 16	11 x 11 x 64	3 x 3 x 16	11 x 11 x 64
Pooling 1	1 x 1 average pool			
Convolution 2	7 x 7 x 32	7 x 7 x 32	7 x 7 x 32	7 x 7 x 32
Pooling 2	2 x 2 max pool			
Convolution 3	11 x 11 x 64	3 x 3 x 16	11 x 11 x 64	3 x 3 x 16
Classification	2 fully-connected, softmax			
Pooling 3			2 x 2 max pool	1 x 1 average pool
Convolution 4			11 x 11 x 64	3 x 3 x 16
Pooling 4			2 x 2 max pool	
Convolution 5			7 x 7 x 32	7 x 7 x 32
Pooling 5			2 x 2 max pool	1 x 1 average pool
Convolution 6			3 x 3 x 16	11 x 11 x 64
Classification			2 fully-connected, softmax	



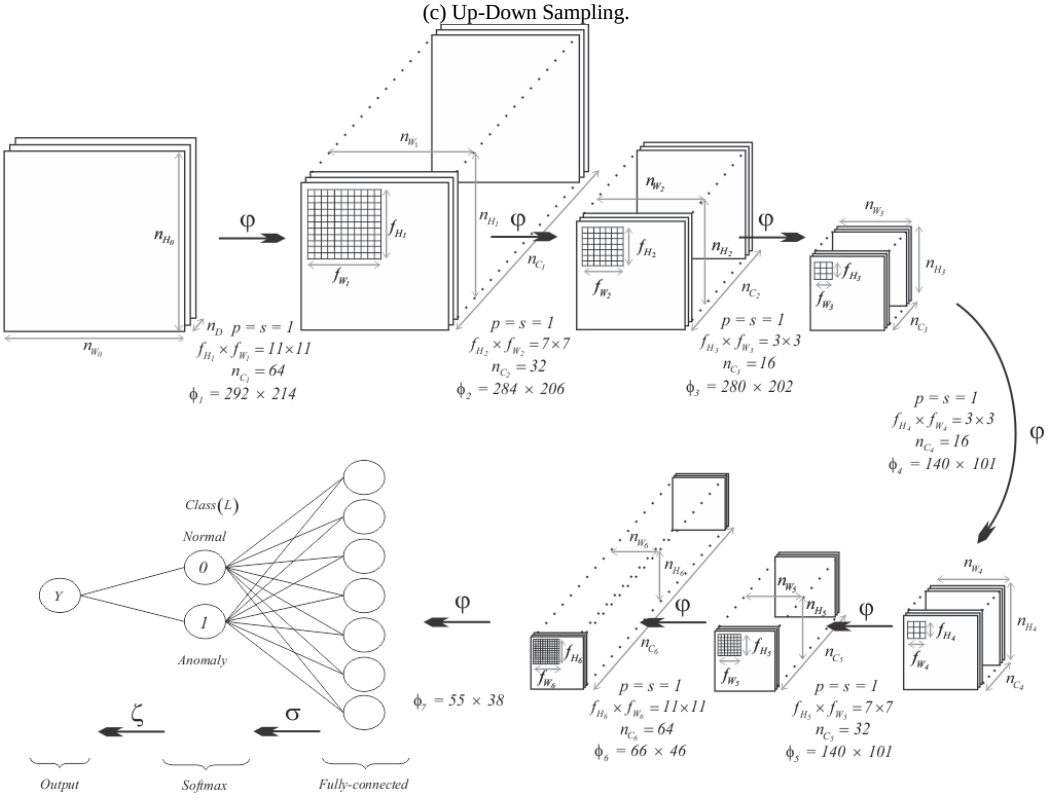
(a) Up Sampling



(b) Down Sampling.



(c) Up-Down Sampling.



(d) Down-Up Sampling.

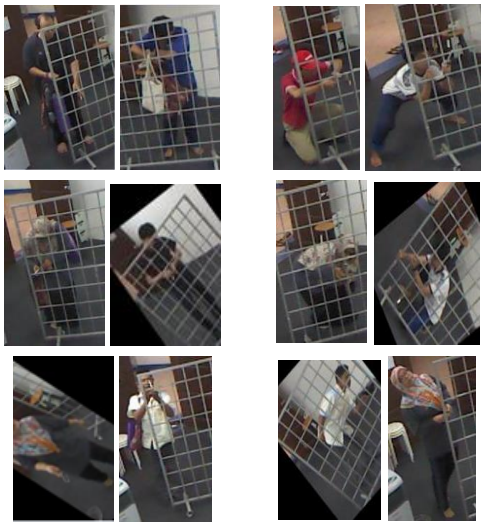
Fig. 3. The Architecture of the Developed CNN Samplings with Hyperparamters and Feature Maps of each Layer.

V. Experimental Results and Discussion

In this study, forensic postures are used, and are referred to as the postures defined by the Royal Malaysia Police (RMP). The definition is interpreted from the Criminal Procedure Code practiced by RMP. The authorities agreed that the observations from their experience are relatively in line with the Malaysian Penal Code. The four most frequent postures recognized during housebreaking crime are squatting, bending, squatting with heels up, squatting with heels down, and kneeling with heels up.

By complying with the requirement of forensic postures, input images that comprised of 9558 color images for each class representing normal and anomaly are collected during data acquisition. Some of the images collected from the footage of participants acted as criminals or otherwise are as in Fig. 4. For each category, 7000 images are randomly selected as the training images whilst the remainder as the testing images.

During experimental, AlexNet and VGG-16 were used in classifying the housebreaking crime postures, but the intention to leverage the architecture of VGG-16 was not achieved since the hardware platform was insufficient to meet the requirement; enormous memory due to the computational cost of 138 million parameters. Thus, four samplings were developed, suitable for low memory platforms to train and test a total of 19116 images together with the AlexNet. The effectiveness of all networks is investigated under two methods. First is the offline test using CCTV videos of housebreaking crime in Malaysia as inputs. Next is the real-time testing using the live feed from a webcam that replicates CCTV as inputs.



(a) (b)

Fig. 4. Input Images, (a) Normal Behaviors, (b) Anomalous Behaviors.

A. Performance of Trained Networks

Referring to Fig. 5, AlexNet with the up-down architecture of convolution layers has the highest training parameters, producing the highest ability to recognize normal behaviors with 99.27% specificity. The developed CNN, Up sampling is the best network for identifying anomalous behaviors according to the sensitivity percentages of 97.26%. Networks with higher training parameters namely AlexNet, Up sampling and Up-Down sampling achieved recognition rate of 97% to 99% in identifying normal and anomalous behavior. These results prove that the number of training parameters contributes to the performance of networks. However, the architectures and hyper parameters are more dominant considering the classification results of all developed samplings are comparable to AlexNet, which have more than 62 times higher computation costs. Here, results showed the ability of AlexNet to identify abnormal and normal behavior based on the sensitivity and specificity obtained.

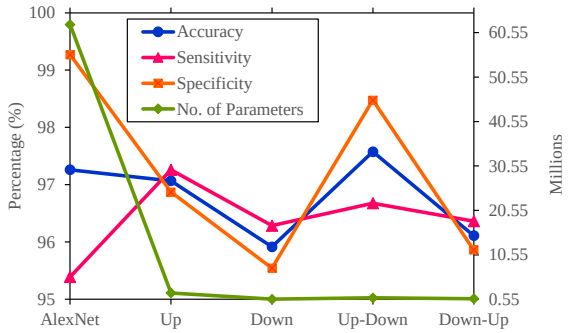


Fig. 5. Performance of Networks in Classifying Normal and Anomalous Behavior.

B. Offline Test

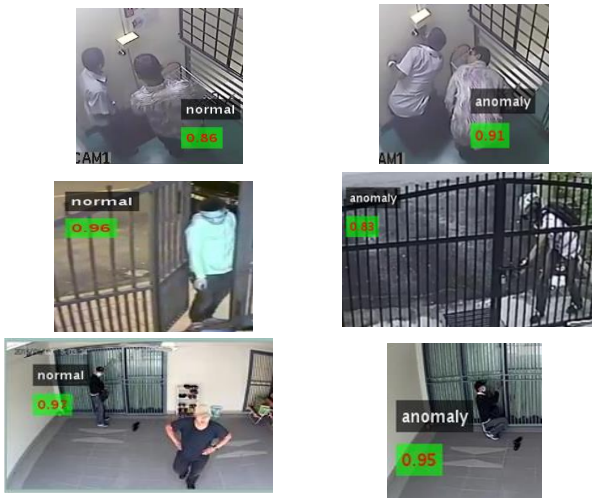
For offline testing, CCTV videos of housebreaking at the gate of residential units in Malaysia were used. The videos recorded single or multiple perpetrators while committing crimes, in broad daylight or at night. The duration of these videos is within 55 seconds to nearly three minutes for housebreaking without using tools. However, videos for housebreaking using tool have a longer video length of two to eight minutes. Each of the perpetrators clearly performs the abnormal characteristics as defined by RMP. Three networks, AlexNet, Up sampling and Down sampling, have successfully recognized both normal and anomalous behaviors from CCTV videos. All networks showed higher ability in identifying normal behaviors than anomalous behaviors following a higher percentage attained by specificity than sensitivity as in Table IV.

TABLE IV. Performance of Trained Networks on Offline Test

Network	Sensitivity (%)	Specificity (%)	Detection Skill
AlexNet	70 to 99	90 to 99	High
Up	80 to 95	80 to 90	Moderate
Down	70 to 80	80 to 90	Moderate
Up-Down	-	100	Failed
Down-Up	-	100	Failed

Detection skill for normal behavior is around 95% to 100% due to the predictable, regular routines at the gate. However, a detection range of 75% to 100% is forecasted for abnormal behavior resulting from the perpetrator's complexity and unpredictable behavior (s). Thus, AlexNet is categorized as 'High' for Detection Skill because it has demonstrated the abilities in detecting as presumed whilst Up Sampling and

Down Sampling as 'Moderate' because lower percentages were recorded for normal and anomalous behavior than the desired results during the classification process. However, moderate detection skill achieved by Down sampling was impressive given its previous performance at the lowest place. Up-Down and Down-Up sampling are categorized as 'Failed' based on the testing from single behavior detection through all videos. Fig. 6 depicts the results during offline test that indicated high accuracy of detection from AlexNet and Up sampling.



(a) (b)

Fig. 6. Offline Detection towards Behaviors at the Gate, (a) Normal Behavior, (b) Anomalous Behavior.

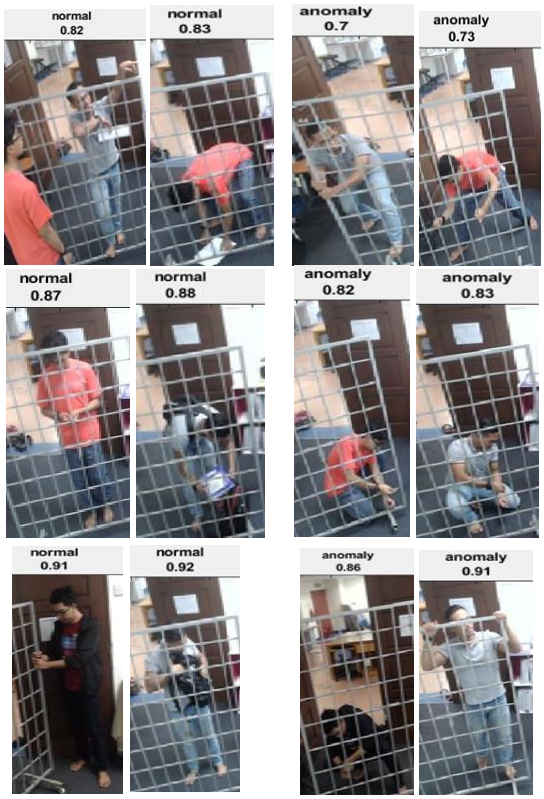
C. Real-Time Test

For the real-time test, live feed images from a webcam were used by the networks to detect within 40 milliseconds to 0.2 seconds. The trial was held in the laboratory and conducted in various situations such as different acts according to the type of gates, for instance, slide gate or push gate, sneaking or lurking, breaking locked gate using a tool for both normal and anomaly behaviors. Participants were required to behave normally at the gate, such as unlocking the padlock or latch and picking up object(s) on the ground, according to their normal habits for routine behavior detection. During anomalous behaviors detection, participants were further requested to impersonate housebreaking crime at the gate according to their interpretation, perspective, and evaluation.

Results showed that AlexNet and Up sampling successfully detecting normal and anomalous behaviors during the real-time test up to 90% recognition rate, as in Fig. 7. Refer to Table V detection achieved by AlexNet is again categorized as 'High' whilst Up Sampling as 'Moderate'. However, for real-time scenarios, Down Sampling is categorized as 'Failed" along with Up-Down and Down-Up sampling since the live feed images are detected as normal throughout the test. Similar results were identified throughout the real-time test, with squatting being highly identified as an anomaly, standing as normal and bending and kneeling were identified more as an anomaly than normal. All the normal activities at the gate were recorded less time duration than anomalous activities, including situations requiring the participants to take out keys from the backpack while holding bags using the other hand.

TABLE V. PERFORMANCE OF TRAINED NETWORKS ON REAL-TIME TEST

Network	Sensitivity (%)	Specificity (%)	Detection Skill
AlexNet	70 to 90	80 to 92	High
Up	70 to 85	85 to 90	Moderate
Down	-	100	Failed
Up-Down	-	100	Failed
Down-Up	-	100	Failed



(a) (b)

Fig. 7. Real-time Detection towards Behaviors at the Gate, (a) Normal Behavior, (b) Anomalous Behavior.

VI. CONCLUSION

In conclusion, the experimental results showed that the architectures, number of parameters, and precisely choosing the hyperparameters are the keys to developing a robust network. It is essential to understand the CNN theoretical and mathematical concepts to develop optimum architecture. The developed samplings proposed in this study are suitable for modest hardware platforms but yielded up to 97% accuracy and succeeded an offline and real-time tests with 97% and 90% recognition rates, respectively. In addition, all samplings were trained using single-subject images for both normal and anomalous behaviors. Yet, the AlexNet and Up sampling could recognize normal and anomalous behaviors for more than one subject and successfully distinguish the anomalous behavior of one person from a group of normal subjects during both offline and real-time tests. The performance of Up sampling has been proven to be on par with the renowned CNN, AlexNet and even more attractive, the computational cost of Up sampling is almost 62 times cheaper.

Finally, CNN has proven its ability in detecting and classify humans based on criminal gait or otherwise. Future work includes developing a classification method to increase the performance and the employment of Faster-RCNN that can enhance its strength in behavior detection. The next stage of work will also explore anomalous human behavior in other potential crime environments, such as banks and high-security areas at airports, warehouses, parking vehicle areas, and car robbery. These initial findings could lead to the development of forensic intelligent surveillance systems that can further help the authorities to decrease the criminal cases rate.

ACKNOWLEDGMENT

This research is funded by the Ministry of Higher Education (MOHE) Malaysia via the Fundamental Research Grant Scheme (FRGS) No: FRGS/1/2019/TK04/UITM/01/3). We would like to thank the Royal Malaysia Police for providing legal information and assisting in developing the forensic gait features and the College of Engineering, UiTM, Selangor, Malaysia, for the facilities and support given in this research.

REFERENCES

[1] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait Verification System for Criminal Investigation," IPSJ Trans. Comput. Vis. Appl., vol. 5, pp. 163–175, 2013, doi: 10.2197/ipsjtcva.5.163.

[2] M. Ben Ayed and M. Abid, "Suspicious behavior detection based on DECOC classifier," 18th Int. Conf. Sci. Tech. Autom. Control Comput. Eng., pp. 594–598, 2017.

[3] L. He, D. Wang, and H. Wang, "Human abnormal action identification method in different scenarios," Proc. 2011 2nd Int. Conf. Digit. Manuf. Autom. ICDMA 2011, pp. 594–597, 2011, doi: 10.1109/ICDMA.2011.148.

[4] W. Lawson and L. Hiatt, "Detecting Anomalous Objects on Mobile Platforms," 2016 IEEE Conf. Comput. Vis. Pattern Recognit. Work., pp. 1426–1433, 2016, doi: 10.1109/CVPRW.2016.179.

[5] N. C. Tay, C. Tee, T. S. Ong, K. O. M. Goh, and P. S. Teh, "A Robust Abnormal Behavior Detection Method Using Convolutional Neural Network," in Computational Science and Technology. Fifth International Conference on Computational Science and Technology. Lecture Notes in Electrical Engineering, vol. 481, R. Alfred, Y. Lim, A. Ibrahim, and P. Anthony, Eds. Singapore: Springer Nature, 2019, pp. 37–47.

[6] A. Al-Dhamari, R. Sudirman, and N. H. Mahmood, "Abnormal Behavior Detection In Automated Surveillance Videos : A Review," J. Theor. Appl. Inf. Technol., vol. 95, no. 19, pp. 5245–5263, 2017.

[7] B. Delgado, K. Tahboub, and E. J. Delp, "Automatic Detection Of Abnormal Human Events On Train Platforms," IEEE Natl. Aerosp. Electron. Conf. (NAECON 2014), pp. 169–173, 2014.

[8] M. Alotaibi and A. Mahmood, "Improved Gait Recognition based on Specialized Deep Convolutional Neural Networks," 2015 IEEE Appl. Imag. Pattern Recognit. Work., pp. 1–7, 2015.

[9] J. M. Shrein, "Fingerprint Classification Using Convolutional Neural Networks and Ridge Orientation Images," IEEE Symp. Ser. Comput. Intell., pp. 1–8, 2017.

[10] W. Zhang and C. Wang, "Application of Convolution Neural Network in Iris Recognition Technology," 2017 4th Int. Conf. Syst. Informatics (ICSAI 2017), pp. 1169–1174, 2017.

[11] S. Dara and P. Tumma, "Feature Extraction By Using Deep Learning: A Survey," 2018 Second Int. Conf. Electron. Commun. Aerosp. Technol., pp. 1795–1801, 2018, [Online]. Available: https://acadpubl.eu/hub/2018-120-6/1/20.pdf.

[12] F. T. George, V. S. P. Patnam, and K. George, "Real-time deep learning based system to detect suspicious non-verbal gestures," I2MTC 2018 - 2018 IEEE Int. Instrum. Meas. Technol. Conf. Discov. New Horizons Instrum. Meas. Proc., pp. 1–6, 2018, doi: 10.1109/I2MTC.2018.8409864.

[13] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 10, pp. 1–20, 2016.

[14] H. Xu, L. Li, M. Fang, and F. Zhang, "Movement Human Actions Recognition Based on Machine Learning," Int. J. Online Biomed. Eng., vol. 14, no. 4, pp. 193–210, 2018.

[15] L. Zhang and P. N. Suganthan, "Visual Tracking with Convolutional Neural Network," 2015 IEEE Int. Conf. Syst. Man, Cybern., pp. 1–6, 2015.

[16] K. Makantasis, A. Doulamis, N. Doulamis, and K. Psychas, "Deep Learning Based Human Behavior Recognition in Industrial Workflows," 2016 IEEE Int. Conf. Image Process., pp. 1–5, 2016.

[17] Y. Pang, S. Syu, Y. Huang, and B. Chen, "An Advanced Deep Framework for Recognition of Distracted Driving Behaviors," 2018 IEEE 7th Glob. Conf. Consum. Electron., pp. 802–803, 2018.

[18] D. H. Hubel and T. N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," J. Physiol., vol. 160, no. 1, pp. 106–154, 1962.

[19] D. H. Hubel and T. N. Wiesel, "Receptive Fields and Functional Architecture of Monkey Striate Cortex," J. Physiol., vol. 195, no. 1, pp. 215–243, 1968.

[20] M. Mathieu, M. Henaff, and Y. Lecun, "Fast Training of Convolutional Networks through FFTs," arXiv:1312.5851v5 [cs.CV], pp. 1–9, 2014.

[21] A. Lavin and S. Gray, "Fast Algorithms for Convolutional Neural Networks," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4013–4021, 2016.

[22] K. Ochiai, N. Toda, and S. Usui, "New Accelerated Learning Algorithm to Reduce the Oscillation of Weights in Multilayered Neural Networks," Int. Jt. Conf. Neural Networks, vol. 1, pp. 914–919, 1992.

[23] S. K. Lenka and A. G. Mohapatra, "Gradient Descent with Momentum Based Neural Network Pattern Classification for the Prediction of Soil Moisture Content in Precision Agriculture," 2015 IEEE Int. Symp. Nanoelectron. Inf. Syst., pp. 63–66, 2015, doi: 10.1109/iNIS.2015.56.

[24] S. Ruder, "An Overview of Gradient Descent Optimization Algorithms," arXiv:1609.04747v2 [cs.LG], pp. 1–14, 2016.

[25] R. Zaheer, "GPU-based Empirical Evaluation of Activation Functions in Convolutional Neural Networks," 2018 2nd Int. Conf. Inven. Syst. Control (ICISC 2018), pp. 769–773, 2018.

[26] M. Wang, "Look-up Table Unit Activation Function for Deep Convolutional Neural Networks," 2018 IEEE Winter Conf. Appl. Comput. Vis., pp. 1225–1233, 2018.

[27] A. A. M. Al-saffar, H. Tao, and M. A. Talab, "Review of Deep Convolution Neural Network in Image Classification," 2017 Int. Conf. Radar, Antenna, Microwave, Electron. Telecommun., pp. 26–31, 2017.

[28] The MathWork, "Softmax layer - MATLAB," 2016. [Online]. Available: https://www.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.softmaxlayer.html. [Accessed: 20-May-2019].

Teacher e-Training and Student e-Learning during the Period of Confinement Caused by Covid-19 in Case of Morocco

Abdessamad El Omari[1]
Laboratory of Physical Chemistry of Materials, Faculty of Sciences Ben M'Sick, Hassan II Universityline Casablanca, Morocco

Malika Tridane[2]
Regional Center of Education and Formation in Professions Boulevard Bir-Anzarane Anfa Casablanca, Morocco

Said Belaaouad[3]
Laboratory of Physical Chemistry of Materials, Faculty of Sciences Ben M'Sick, Hassan II Universityline Casablanca, Morocco

***Abstract*—The rapid advance of the new coronavirus imposes several drastic measures on the authorities to contain the virus and prevent its spread. The Ministry of Education has decreed the suspension of face-to-face classes in all schools. The pedagogical continuity, now a priority, must be carried out remotely through video conferencing, platforms, video capsules The main questions of this research are: What are the success factors of distance learning and training in Moroccan context? This raises a whole range of questions: are teachers able to adopt this new teaching method? Are the means available and adequate to ensure distance learning? What about the training of teachers to cope with this unexpected radical change? Based on the results obtained from a population of 126 teachers, we discovery that 91% of teachers said that they have adopted the online teaching but it is not really e-learning as recognized by the specialists, its objective is rather to maintain communication with the students. While 9% have not used this mode of teaching, they point out that this type of teaching does not guarantee equal opportunities for learners. We have therefore concluded that the necessary material resources must be made available to ensure the success of this type of teaching, such as computers and the Internet, as well as the necessary training for teachers to develop their skills associated with managing distance learning.**

***Keywords*—COVID-19; e-learning; e-training; change; innovation**

I. Introduction

Morocco, like other countries in the world, experienced a crisis due to the spread of the COVID-19 epidemic in early March 2020.

In order to control this situation, the government has taken a series of decisions by stopping air traffic in order to isolate the country and control internal and incoming infected cases before stopping the movement of airports.

It also closed all places that could contribute to the spread of this epidemic. With the increasing number of infections announced by official communications from the Ministry of Health, the decision of containment has been taken, preventing citizens from leaving their homes except for the most urgent needs.

These prevention measures affected all vital areas; the education and training sector was no exception. The Ministry of National Education, Vocational Training, Higher Education and Scientific Research (MNEVTHESR) suspended face-to-face classes in order to prevent the spread of this epidemic in schools [1].

To ensure pedagogical continuity, the MNEVTHESR launched a communiqué encouraging teachers to adopt distance learning [2], which is not obvious since there was no prior preparation to guarantee the success of this change.

Problematic: The adoption of distance learning leads us to ask this main question: What are the success factors of distance learning and training in Moroccan context? This raises a whole range of questions: How were teachers able to deal with this situation that arose? Did they receive appropriate support to migrate to e-learning? Did they engage in this new teaching modality? How satisfied are they? And what is their attitude towards e-training and e-learning?

The answer to all these questions will help us to achieve the objective of our study.

Purpose of the study: The main aim of this study is to define factors for success of e-training and e-learning, while determining the causes behind the limited use of these modes of training and teaching, whether teachers engage in e-learning, their attitudes towards this type of training and their feelings about his impact on their teaching practices and determining the constraints of such training.

II. Theoretical Background

New technologies for processing, sharing and using information are emerging day by day. The field of learning is also beginning to change its approach and method with the integration of these new technologies, first, by complementing the classical models, by integrating these technologies into classroom practices, and why not replace them later, by adopting online teaching [3].

The subject of the adoption of information and communication technologies in education and especially in training and e-learning, has given rise to a number of theories, concepts and standards, which aim at understanding and analyzing the factors that influence the acceptability and implementation of this technological innovation.

Among these theories are Fishbein and Ajzen's theories of reasoned action and planned behavior [4], Davis's Model of Technology Acceptance (MTA) [5], Brangier's symbiotic model [6], and Ram's theory of resistance to innovation [7]. These theories might help us to understand the factors that can impact the success of e-learning and e-training.

The narrow scope of this research does not allow us to address them in detail. On the other hand, the presentation of two fundamental concepts is necessary: e-training and e-learning.

A. e-Learning

Oxford dictionary define e-learning as "a system of learning that uses electronic media, typically over the internet". Bowles and al. considers it as "encompasses any type of learning content that is delivered electronically" [8]. Hoppe and al. considers e-learning as "a learning supported by digital electronic tools and media" [9]. All these definitions focus on the idea of using electronic tools and media, but there are other dimensions to this teaching practice. for Abrami and al. e-learning is "the development of knowledge and skills through the use of information and communication technologies (ICTs), particularly to support interactions for learning – interactions with content, with learning activities and tools, and with other people" [10].

This definition underlines the interaction aspects of the e-learning process. The student is therefore not only a passive receiver.

B. e-Training

Before defining e-training we must present a definition of training. McClelland defined it as "an activity that changes people's behaviors in an organization. Increased productivity is meant to be the most important reason for training" [11].

E-training is defined as "Training delivered through the electronic means, which could be Web-based training programs and activities" [12]. It is also defined as "a separation of trainer from trainee and part of teaching and training through instruction, observations, or processes focused on providing needed skills and knowledge to meet immediate business goals" [13].

So, the e-training concept in our study is defined as using ICT to develop teachers' knowledge and skills and change their behaviors related to teaching practices in dedicated online teaching environments.

C. Challenges of e-training and e-learning

The operationalization of these concepts has given rise to reflection on issues and challenges of the implementation of online teaching and training practices. Mueen Mohsin and al listed six main issues and challenges which are: lack of awareness, low adoption rate, bandwidth issue and connectivity, language barrier, difficult in engaging learners online, and lastly computer literacy and digital divide [14]. These challenges will increase especially with the conditions of the Covid 19 pandemic.

III. Materials and Methods

A. The Sample

The empirical part of this study is based on a survey of 126 teachers from different regions and different levels (primary, secondary college and qualifying) during the confinement period from March 16 to June 30, 2020. The selection of the sample was done in a simple random way, as we randomly selected the people to answer our survey. The proportion is 42 women and 84 men. The other characteristics are presented in the following Table I as follows.

TABLE I. Presents the age of the Sample

Genre	Age					
	<36 years		36y to 45y		> 45 years	
Men	38	30%	28	22%	18	14%
Women	19	15%	16	13%	7	06%
Total	57	45%	44	35%	25	20%

Table II shows the seniority of the sample.

TABLE II. Seniority of the Sample

Genre	Seniority					
	<11years		11y to 15y		>15years	
Men	45	36%	10	08%	29	23%
Women	26	21%	2	02%	14	10%
Total	71	57%	12	10%	43	43%

Table III shows the location of the establishment of our sample.

TABLE III. Environment of the Establishment

Genre	Environment of the establishment					
	Urban		Periurban		Rural	
Men	40	32%	18	14%	26	21%
Women	29	23%	9	07%	4	03%
Total	69	55%	27	21%	30	24%

Table IV shows the teaching level of our sample.

TABLE IV. Teaching Level

Genre	Teaching level					
	primary		Middle school		Highschool	
Men	46	37%	11	09%	27	21%
Women	23	18%	05	04%	14	11%
Total	69	55%	16	13%	41	32%

B. Data Collection Tool

The data collection tool used to conduct this study is an online survey that is composed of three parts:

- The first one contains general information about the sample.

- The second one is about teacher e-training:
  - The existence of e-training courses for teachers;
  - Participation in e-training courses;
  - Reasons for not taking advantage of e-training courses;
  - Avenues for the development of distance e-training for teachers;
- The third concerns the adoption of e-learning by teachers:
  - The performance of teachers during the e-learning process and their degree of satisfaction;
  - Reasons why teachers have not all adopted e-learning.

IV. Results and Discussion

In this part, we will propose an analysis of the responses of our sample according to two main axes: e-training and e-learning.

A. e-Training

At first, we will see if the teachers have an idea about the existence of distance training for teachers on the internet. 52% say there are online training courses. While 48% have no idea about the offer of online training. However, half of 52% did not receive any training during this period of confinement. This means that only 26% of all sample have benefited from online training but 74% of teachers have not benefited from online training. This prompts us to look for the causes behind this situation, otherwise why teachers have not taken online training? The results are presented as follow in Table V.

TABLE V. Reasons for Not Following e-Training Courses

Reasons for not following e-training courses	Frequency
I did not find these courses	38%
I was not asked to participate	39%
I didn't look for these courses	23%
I don't need this training	05%
I find that I don't have the necessary skills to follow these courses.	15%
My prerequisites are insufficient to follow these courses.	37%

For those teachers who have attended e-training courses, has it had an impact on their practices? We have proposed this question as a Likert scale, it quantifies the qualitative information may result in a score. The results are presented as follow in Fig. 1.

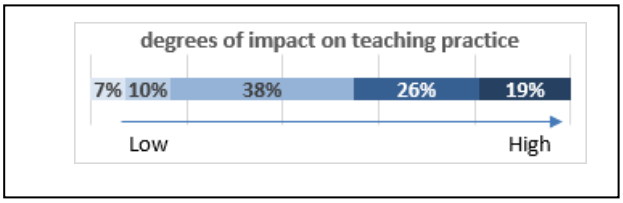


Fig. 1. Degree of Impact on Teaching Practice.

Almost half of interviewees 45% feel significant impact of training on teaching practices. The other half, on the other hand, see little impact.

These data prompted us to ask a legitimate question: why do e-training courses have a limited impact on improving teaching practices?

The results are shown in Fig. 2.

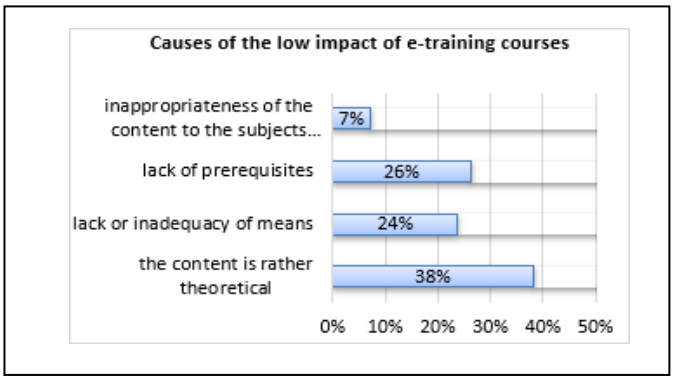


Fig. 2. Causes of the Low Impact of e-training Courses.

A simple reading of the answers to these questions reflects the need to review the engineering of these training courses and relate them to the actual needs of teachers. The introduction of e-training must be accompanied by a paradigm shift in attitudes, pedagogical concepts and instructional engineering practices [15]. It must also promote a better balance between theoretical knowledge and those acquired in practical environment [16].

In order to have a positive impact of the training courses, we need to know what kind of form do teachers prefer?

The results are presented in the Table VI.

TABLE VI. Mode of Training

Mode of training	Frequency
Hybrid Training	57%
Online training	11%
Analysis of practices	39%
Face-to-face training	48%
Training through coaching	30%

From the percentages obtained it seems clear that teachers are not very interested to online learning. On the other hand, they are rather inclined to other forms of training.

B. e-Learning

The lack of teacher training and support has had an impact on the adoption of e-learning and the quality of this modality.

For the adoption of e-learning with students during confinement, the results are as follow in Fig. 3.

Only 9% have not used this mode of teaching, while 91% of teachers said that they have adopted it. But can we talk really about e-learning where the teacher has a platform dedicated to this type of teaching?

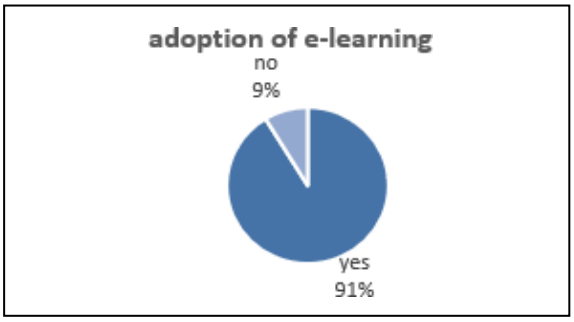


Fig. 3. Levels of Teacher's Satisfaction.

Understanding the quality of this type of teaching has led us to ask two questions; the first concerns the means used and the second concerns measuring the degree of satisfaction of teachers.

The answer of the first question is shown in Table VII.

TABLE VII. THE MEANS ADOPTED DURING E-LEARNING

The means adopted	Frequency
Social networks	88%
Videoconferencing tools	13%
Microsoft-Teams platform recommended by the Ministry	22%

The most widely used means are social networks, this shows that the main aim is to maintain communication with students more than e-learning itself. Just 22% of teachers have used Microsoft-teams platform for this mode of teaching.

To measure the degree of satisfaction we used a question with the Likert scale. The results are presented in Fig. 4.

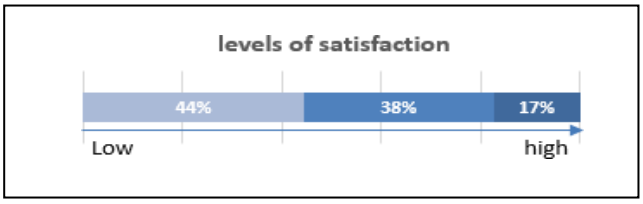


Fig. 4. Adoption of e-learning.

Only 17% of teachers expressed satisfaction. While 44% are dissatisfied.

The reasons for dissatisfaction for those who have adopted this modality are presented in Table VIII.

TABLE VIII. REASONS FOR TEACHER DISSATISFACTION

Reasons for dissatisfaction	Frequency
Unsatisfactory capacity of teachers to use the platforms.	12%
Unsatisfactory ability of students to use the platforms.	37%
The low quality of the means and tools used by teachers.	36%
The low quality of the means and tools used by students.	54%
Not all students have access to e-learning.	81%

Even if Morocco has mobilized since 2004 to set up a national strategy for integrating information and communication technologies into the educational process [17]. we see that there is still a lack of access to technological tools either for teachers or for students. There are hidden digital differences that impede equal access to digital resources for all learners [18]. There is a lack of home computer access 61.5%, lack of computer training 70.4% and Absence of oriented use (research activities, exercises, courses, etc.) [19].

Almost all of teachers who did not use e-learning argue that the lack of equitable access to e-learning for learners makes its implementation practically impossible.

V. CONCLUSION

This study shows us that most of our population (91%) adopts e-learning, but it is not really e-learning as the specialists recognize, it is rather aimed at maintaining communication with the students. While a minority (9%) has not adopted this mode of teaching, their pretext is that this type of teaching does not guarantee equal opportunities for learners.

We have concluded that the success of this type of teaching implies the provision of the necessary material resources, such as computers and the Internet, to teachers and students, as well as the necessary training for teachers to develop their skills related to the management of e-learning.

VI. RECOMMENDATIONS

The success of distance learning and training in the Moroccan context requires a change in teacher's representations and perceptions of these modes of work. This change must be accompanied by responsible for education and training, with a view to ensuring the optimum conditions, namely: a careful analysis of teacher's needs in order to design quality training arrangements focusing more on the practical side and the provision of teachers and students with equipment and resources that can improve this teaching and training method.

Several digital and hybrid methods can be taken into consideration which makes it possible to improve the quality of teaching-learning in the context of online training.

We can use the work of Daaif in 2019 that developed applications for the simulation of practical work in geometric crystallography as well as other applications for modeling chemical reactions by the Monte-Carlo method in kinetics [20-21-22].

VII. LIMITATIONS

The limitations of this study include:

- The limited number of the target population which does not allow for generalizing the results of this research.
- Also, the online questionnaires do not guarantee the credibility of the answers of the members of the target population, but the studies conducted in this field (in Moroccan context) confirm the results obtained by the present research.

REFERENCES

[1] Ministerial announcement concerning the interruption of face-to-face courses and their replacement by distance courses.

[2] Guidance Note on the Response of Education Systems to COVID19. March 25, 2020.

[3] J. Muftisada, J. Daaif, M. Tridane S. Belaaouad "Use of smartphones in learning and techno-pedagogical integration of artificial intelligence tools as a prospect for intelligent learning: Case of Moroccan students from Hassan II University of Casablanca" Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter252892020.pdf. https://doi.org/10.30534/ijeter/2020/252892020 ISSN 2347 – 3983, International Journal of Emerging Trends in Engineering Research Volume 8. No. 9, 6496-6504, September 2020.

[4] Ajzen. I., Fieshbein. M., (1975), "Belief, attitude, intention and behavior ", edition Addison Wesley.

[5] Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, 13(3), 319. doi:10.2307/249008.

[6] Éric Brangier, Aude Dufresne, Sonia Hammes-Adelé " Symbiotic approach to the human-technology relationship: perspectives for computer ergonomics " Le Travail Humain, tome 72, no 4/2009, 333-353.

[7] Ram Sudha, "A model of innovation resistance ", In Advances in Consumer Research, n° 14,1987, p. 208-212.

[8] Bowles, M. 2004. Relearning to e-learn: Strategies for electronic learning and knowledge. Melbourne University Publishing.

[9] H.U. Hoppe , R. Joiner, M. Milrad & M. Sharples, Guest editorial: Wireless and Mobile Technologies in Education, Journal of Computer Assisted Learning (2003) 19, 255-259".

[10] Abrami, P., Bernard, R., Wade, A, Borokhovski, E., Tamin, R., Surkes, M., and Zhang, D. (2008). A review of e-learning in Canada: Rejoinder to commentaries. Canadian Journal of Learning and Technology, 32(3), p. 30.

[11] McClelland, S.D. 2002. A training needs assessment for the united way of dunn county wisconsin. University of Wisconsin.

[12] https://www.igi-global.com/dictionary/does-learning-improve-communication-among/8945.

[13] Echard, R.D. and Z.L. Berge, 2008. Quality management builds solid e-training. Administering examinations for quality control in distance education: The National Open University of Nigeria Perspective EC IBARA.

[14] Mueen Mohsin , Rosnafisah Sulaiman , a study on e-training adoption for higher learning institutions, International Conference on Teaching and Learning in Education, 2013, International Journal of Asian Social Science, 2013, 3(9):2006-2018.

[15] M.Bassiri, M.Radid, and S. Belaaouad "Modeling of the Teaching–Learning Process in E-Learning" K29959_C002.indd, Shaping the Future of ICT, 07-06-2017, Page 15-40, (2017).

[16] A.El yadari, M.Tridane, M.Radid ,S.Belaaouad "Engineering of a training program for future teachers based on information and communication technologies" International Journal of Advanced Trends in Computer Science and Engineering. http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4681.42019.pdf https://doi.org/10.30534/ijatcse/2019/4681.42019.

[17] Mazouak A., " Modeling of Digital Media in The Management of Educational Performance in Morocco School's" Iraqi Journal of Science, 2021, Special Issue, pp: 17-23 DOI: 10.24996/ijs.2021.SI.1.3.

[18] O.Dardary, J.Daaif, M.Tridane, S.Belaaouad "Distance learning in the age of covid – 19: between perspective and reality " International Journal of Engineering Applied Sciences and Technology, Published Online September 2020 in IJEAST (http://www.ijeast.com) Vol. 5, Issue 5, ISSN No. 2455-2143, Pages 46-52, 2020.

[19] O.Dardary, Z.Azar, M.Tridane, S. Belaaouad "Engineering of a Training Device and Skills Through the Integration of New Information and Communication Technologies in the Field of Exact Sciences" International Journal of Advanced Trends in Computer Science and Engineering http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4581.42019.pdf https://doi.org/10.30534/ijatcse/2019/4581.42019.

[20] Daaif J., Zerraf S., Tridane M., Benmokhtar S., Belaaouad S. (2019a). Pedagogical engineering to the teaching of the practical experiments of chemistry: Development of an application of three-dimensional digital modelling of crystalline structures. Cogent Education, 2019, 6(1), 1708651. DOI: https://doi.org/10.1080/2331186X.2019.1708651.

[21] J. Daaif, S. Zerraf, M. Tridane, S. Benmokhtar, S. Belaaouad. (2019b). Technological Innovation in Teaching and Research in Chemical Science: Development of a Computer Application for the Simulation of the Practical Works of Crystallography. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3.

[22] Daaif, J., Zerraf, S., Tridane, M., El Mahi Chbihi, M., Moutaabbid, M., Benmokhtar, S., Belaaouad, S. (2019c). Computer simulations as a complementary educational tool in practical work: Application of monte- carlo simulation to estimate the kinetic parameters for chemical reactions. International Journal of Advanced Trends in Computer Science and Engineering, 8(1.4 S1), pp. 249–254. https://doi.org/10.30534/ijatcse/2019/3881.42019.

Performance Evaluation of Different Raspberry Pi Models for a Broad Spectrum of Interests

Eric Gamess
MCIS Department
Jacksonville State University
Jacksonville, Alabama, USA

Sergio Hernandez
Information Security
Citibank, New York
USA

***Abstract*—Now-a-days, Single Board Computers (SBCs), especially Raspberry Pi (RPi) devices, are extensively used due to their low cost, efficient use of energy, and successful implementation in a wide range of applications; therefore, evaluating their performance is critical to better understand the applicability of RPis to solve problems in different areas of knowledge. This paper describes a comparative and experimental study regarding the performance of five different models of the RPi family (RPi Zero W, RPi Zero 2 W, RPi 3B, RPi 3B+, and RPi 4B) in several scenarios and with different configurations. To conduct our multiple experiments on RPis, we used a self-developed and other existing open-source benchmarking tools allowing us to perform tests that mimic real-world needs, assessing important factors including CPU frequency and temperature during stressful activities, processor performance when executing CPU-intensive processes such as audio and file compressions as well as cryptographic operations, memory and microSD storage performance when executing read and write operations, TCP throughput in different WiFi bands, and TCP latency to send a specific payload from a source to a destination. Our experimental results showed that the RPi 4B significantly outperformed the other SBCs tested. In addition, our research indicated that the RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ had similar performance. Finally, the RPi Zero 2 W showed a much higher capacity than its predecessor, the RPi Zero W, and seems to be a perfect replacement when upgrading, since they have the same form factor and are physically interchangeable. With this study, we aim to guide researchers and hobbyists in selecting adequate RPis for their projects.**

***Keywords*—Performance evaluation; benchmarks; raspberry pi; single board computer**

I. Introduction

One of the most popular Single Board Computer (SBC) is the Raspberry Pi (RPi), with a vast online community of users around the world. Build on open-source principles and motivated by the non-profit incentive to increase global access to computing and solve a variety of real-world challenges using digital technology, these low-cost SBCs bring together external hardware, sensors and controller interfaces, with user-friendly programming capabilities, high connectivity, and desktop functionality [1].

RPis are being employed in a broad range of projects across diverse topics and research fields, including the Internet of Things (IoT) that has become widely used in recent years in an extensive range of applications from smart cities and industries to water monitoring [2][3]. Examples of successful uses of RPis can be found in the field of Artificial Intelligence (AI) and Machine Learning (ML), where researchers have been highlighting not only a good performance but also a low energy consumption [4][5]. Additionally, the usage of RPis has spread to other relevant areas such as cybersecurity, energy, health, education, to name a few [6].

Considering the frequent usage of RPis in applications, it is essential to deeply analyze how they behave and perform under different conditions for a given period of time, to better understand their capabilities and limitations. In this paper, we carried out a comparative and experimental study of five different models of the RPi family (RPi Zero W, RPi Zero 2 W, RPi 3B, RPi 3B+, and RPi 4B). To do so, we conducted several experiments by using a benchmarking tool that we developed as well as existing open-source benchmarking tools to evaluate the performance of RPis in terms of processor frequency and temperature, CPU use level, RAM and microSD access performance, TCP throughput and latency, among others. We think the results of this study might guide scientists and hobbyists in selecting adequate models of RPi for their projects, according to their budget and performance requirements.

The rest of the paper is structured as follows. Section II discusses the related work. The description of the testbed environment is done in Section III. Section IV reports and discusses the results of our performance evaluation of the different RPi models. Finally, Section V concludes the paper and gives directions for future work.

II. Related Work

Due to the popularity and constant evolution of the Raspberry Pi models, several works about the performance evaluation of these devices have been performed. We reviewed the literature in order to examine and understand areas, methods, and available tools to assess their performance.

Morabito [7] performed an assessment of container-based technologies running on top of a Raspberry Pi 2 Model B using different types of workloads to challenge a specific subsystem of the hardware under test. This study aimed to evaluate the use of Docker containers in constrained environments, providing a detailed performance analysis. Experiments in a testbed were conducted, and metrics such as CPU execution time, power consumption, memory speed, network bandwidth, and protocol overhead for MySQL and Apache were reported.

Other works [8][9] examined the performance of Intrusion Detection Systems (IDS) running on RPis. Kyaw, Chen, and Joseph [8] presented the results of an experiment comparing two open-source IDSs (Snort and Bro) on a Raspberry Pi 2 Model B, with the main goal of determining their performance, efficiency, and efficacy for use in computer network environments, where cost is a determining factor. On the other hand, Aspernäs and Simonsson [9] compared two RPis (Raspberry Pi 1 Model B+ and Raspberry Pi 2 Model B) while examining the traffic for intrusion detection in a home environment.

The authors of [10] carried out a performance analysis of SNMP [11] agents running in three different RPis (Raspberry Pi Zero W, Raspberry Pi 3 Model B, and Raspberry Pi 3 Model B+). Numerous experiments varying different parameters, requests, versions, and security models of SNMP were performed.

Another related work was consulted in [12], where a performance evaluation of RESTful frameworks on two Raspberry Pi 1 Model B was carried out. In this study, experiments involving combinations of factors such as device CPU frequency and message size with different configurations or levels were conducted by comparing two web services frameworks (Axis2 and CXF) to understand better not only the energy consumption, but also the relationship between performance and energy consumption in RPis.

Guamán, Ninahualpa, Salazar, and Guarda [13] did a comparative study between the MQTT [14][15] and the CoAP [16][17] protocols for IoT in an IEEE 802.11 environments, using a Raspberry Pi 3 Model B+. For the analysis of network parameters, traffic injection tests were carried out with specific tools, using different bandwidths and data sizes. Another work in this direction was done in [3], where the authors presented performance measurements of the Raspberry Pi Zero W working as an IoT gateway between local sensors and a public MQTT [14][15] broker running in the cloud. The experimental results demonstrated its performance using the following metrics: CPU utilization, temperature, as well as the rate of received MQTT messages under different levels of Quality of Service (QoS).

Other studies in the area are focused on benchmarking cryptographic algorithms on RPis. For instance, Hawthorne, Kapralos, Blaine, and Matthews [18] compared the performance of three computing systems for three leading cryptographic algorithms (AES, Twofish, and Serpent). The three computing systems considered were: (1) a cluster of Raspberry Pi 3 Model B+, (2) a power-efficient next unit of computing (NUC), and (3) a mid-range desktop (MRD). The metrics reported by this work included encryption/decryption throughput and power consumption. Similarly, the study in [19] aims to analyze the performance of symmetric encryption algorithms (DES and AES) within the PHP programming language using RPis. The authors reported parameters such as the time and memory consumption to encrypt/decrypt messages.

Although multiple studies on performance evaluation have been performed, they mainly analyze how specific technologies (e.g., containers) behave when run on certain RPi models. Thus, parameters that were evaluated are intrinsically related to a particular environment and do not allow a broad understanding of how the main RPi subsystems respond to different conditions, based on CPU workload, I/O requirement, network traffic, amongst others. Since the potential of the RPi has not been broadly analyzed, there is a gap for those who want to have a more general view of the performance of an RPi, for scenarios that were not covered by the actual specialized literature.

Unlike previous works, our paper considers five different models of RPis, and our assessment covers a broad spectrum of interests. Moreover, the Raspberry Pi Zero 2 W was released at the end of October 2021, and no other scientific work to evaluate its performance was found at the time of performing this study. Therefore, with this paper, we aim to guide scientists and hobbyists in their selection of an adequate model of Raspberry Pi according to their budget and performance requirements.

III. Description of the Testbed Environment

A. Models of Raspberry Pi used in the Experiments

For our assessment, we had the following Raspberry Pi SBCs: two Raspberry Pi Zero W, two Raspberry Pi Zero 2 W, two Raspberry Pi 3 Model B, one Raspberry Pi 3 Model B+, and one Raspberry Pi 4 Model B (8 GB of RAM). Some of their technical specifications are presented next:

- Raspberry Pi Zero W (RPi Zero W): It is based on a 32-bit Broadcom BCM2835 single-core ARM1176JZF-S SoC @ 1.0 GHz, 512 MB of RAM, one 2.4 GHz IEEE 802.11b/g/n WiFi interface, one micro USB On-The-Go port, one mini HDMI connector, and one microSD card slot [20].
- Raspberry Pi Zero 2 W (RPi Zero 2 W): It is based on an RP3A0-AU, which consists of the integration of a 64-bit Broadcom BCM2710A1 quad-core Cortex-A53 @ 1.0 GHz and 512 MB of RAM, in a single chip. It also has one 2.4 GHz IEEE 802.11b/g/n WiFi interface, one micro USB On-The-Go port, one mini HDMI connector, and one microSD card slot [21]. It can be easily overclocked to 1.3 GHz with an adequate heat sink.
- Raspberry Pi 3 Model B (RPi 3B): It is based on a 64-bit Broadcom BCM2837 quad-core Cortex-A53 SoC @ 1.2 GHz, 1 GB of RAM, one 10/100 Mbps Ethernet interface, one 2.4 GHz IEEE 802.11b/g/n WiFi interface, four USB 2.0 ports, one full-size HDMI connector, and one microSD card slot [22].
- Raspberry Pi 3 Model B+ (RPi 3B+): It is based on a 64-bit Broadcom BCM2837B0 quad-core Cortex-A53 SoC @ 1.4 GHz, 1 GB of RAM, one Gigabit Ethernet interface over USB 2.0 (maximum throughput 300 Mbps), one dual-band 2.4 GHz and 5 GHz IEEE 802.11b/g/n/ac WiFi interface, four USB 2.0 ports, one full-size HDMI connector, and one microSD card slot [23].

- Raspberry Pi 4 Model B (RPi 4B): It is based on a 64-bit Broadcom BCM2711 quad-core Cortex-A72 SoC @ 1.8 GHz, 1/2/4/8 GB of RAM, one Gigabit Ethernet interface, one dual-band 2.4 GHz and 5 GHz IEEE 802.11b/g/n/ac WiFi interface, two USB 2.0 ports, two USB 3.0 ports, two micro-HDMI connectors, and one microSD card slot [24].

In all the SBCs, we inserted a 64 GB SanDisk Extreme microSDXC UHS-I Memory Card (SDSQXA2-064G-GN6MA) with the operating system preinstalled. It is considered as one of the fastest microSD cards of the market, with up to 160 MB/s and 60 MB/s for the reading and writing speeds, respectively. Also, unless otherwise stated, all the experiments were carried out with no cooling solution for the RPi Zero W, RPi Zero 2 W, RPi 3B, and RPi 3B+. However, for most of the experiments, we chose to control the temperature of the RPi 4B with a small fan since its temperature can dramatically increase, in contrast to the other RPis that we selected for our study.

B. Operating Systems for Raspberry Pi

Several operating systems are available for Raspberry Pi. We opted for the most popular one developed by the Raspberry Pi Foundation, and known as Raspberry Pi OS (32-bit version). It is based on Debian Bullseye. The last version was released in October 2021. Three options of this operating system are available: (1) Raspberry Pi OS Lite, (2) Raspberry Pi OS with Desktop, and (3) Raspberry Pi OS with Desktop and Recommended Software. The "Lite" option is a minimal image consisting of 493 packages without an X-window manager. Hence, it runs fast and is more suitable for a server environment. The "Desktop" option is a superset of the "Lite" option, with a total of 1324 packages. It is more oriented to end-users since it includes Openbox as the window manager and LXDE as the desktop environment. The "Desktop and Recommended Software" option consists of 1944 packages and has all the "Desktop" option features, but also includes an office productivity suite (LibreOffice) and additional supports for developers (Erlang, Node.js, Ruby, Java, Wolfram, Apache Ant, BlueJ, Firebird, and Greenfoot). We chose the "Lite" option for all our experiments since our performance evaluation is more targeted towards a server installation in headless mode.

There is a beta version of Raspberry Pi OS for 64-bit architecture. Hence, it does not work with the RPi Zero W since it is equipped with a 32-bit processor. The Raspberry Pi Foundation released its last version in October 2021, and we could not make it works with the RPi Zero 2 W.

C. Testbed Used for the Network Experiments

We assessed the performance of the networking system of the different RPis. To do so, we utilized the testbed of Fig. 1. It consisted of two network devices connected through a wireless router. In some experiments, the network devices were two identical RPis. In other experiments, one network device was an RPi, while the other one was a PC. The two network devices were placed 4 meters from the wireless router, with no obstacles between them. For the wireless router, we used a NETGEAR AC1200 Smart WiFi Router R6220. It had the following characteristics: an 880 MHz MediaTek processor width two radio bands (IEEE 802.11b/g/n in the 2.4 GHz band and IEEE 802.11a/n/ac in the 5 GHz band), 128 MB of flash, 128 MB of RAM, and five 10/100/1000 Mbps Ethernet ports (one WAN and four LANs).

The PC had the following specifications: Dell OptiPlex 3030 AIO, with a 64-bit Intel quad-core i3-4130 CPU at 3.4 GHz, 16 GB of RAM, a 512 GB SSD, a 1 Gbps Ethernet NIC, and an Intel Wireless 7260 Network Adapter (dual-band WiFi adapter with support to IEEE 802.11a/b/g/n/ac). Debian amd64 11.1 (codename "Bullseye") was installed as the operating system.

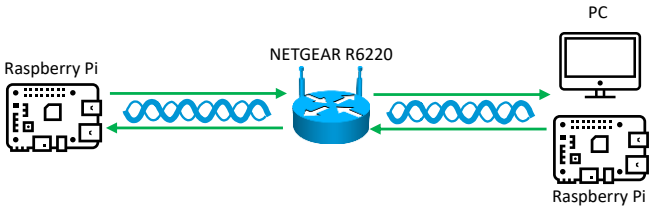


Fig. 1. Testbed for the Network Experiments.

IV. Performance Results and Analysis

In this section, we made a performance evaluation of the considered RPis (RPi Zero W, RPi Zero 2 W, RPi 3B, RPi 3B+, and RPi 4B) in several scenarios, reporting different parameters. Each experiment was executed several times, and the reported results are an average of the repeated experimental runs. By repeating and averaging, we ensure the consistency of our empirical findings.

A. Evaluation of the Temperature with Stressberry Test

The processor of an RPi can overheat if it does not have enough cooling. This overheating problem is more frequent in powerful CPUs, such as the one of the RPi 4B. When necessary, CPU throttling (also known as dynamic frequency scaling) will occur to decrease the electrical energy being consumed, and in turn, to reduce the heat generation. Small heat sinks, specific cases, fans, and other solutions can be utilized to cool down the CPU of an RPi.

The Stressberry test [25] can be used to reveal if the CPU of an RPi can run at maximum load in its case/environment without overheating, and being forced to slow down. It has four phases. In the first phase, Stressberry waits until having a stable temperature. To do so, it lets the CPU idle and just takes a temperature sample every 60 seconds. This initial phase ends when two consecutive samples (previous and current temperatures) are the same. It is noted that no temperature and frequency samples are stored in the resulting file during the first phase. In the second phase, the test lets the CPU idle for 150 seconds, storing samples of the base temperature and base frequency in the resulting file every two seconds. Then, in the third phase, all the cores of the CPU are stressed with a maximum load for 300 seconds. The variation of the temperature and frequency are also saved in the resulting file during this phase. In the final phase (fourth phase), the test lets the CPU idle for another 150 seconds, to have an insight on how fast it can cool down. Here also, the variation of the temperature and frequency are recorded in the resulting file. The entire process takes 600 seconds (10 minutes). When the

test ends, the resulting data (temperature and frequency) that were stored in the resulting file can be processed and plotted.

Fig. 2 shows the instructions that we used to install and run Stressberry. Python 3.x is required, and its version was verified with Line 01. Lines 02 and 03 allow the installation of the dependencies and Stressberry, respectively. The version of Stressberry was obtained in Line 04 (in our case, version 0.3.3), while Line 05 displays some help about the software usage. Stressberry was run with the instruction of Line 06, where option `-i` specifies the idle time in seconds before (phase 2) and after (phase 4) the stress portion (phase 3), while option `-d` indicates the stress test duration in seconds (phase 3). Line 07 generates a graphical representation of the experiment for the temperature, while Line 08 does it for both the temperature and frequency.

```
01: python --version
02: apt-get install stress python3-pip libatlas-base-dev \
 libopenjp2-7-dev
03: pip3 install stressberry
04: stressberry-run -v
05: stressberry-run -h
06: stressberry-run -i 150 -d 300 resFile.dat
07: stressberry-plot resFile.dat -o temp.png --not-transparent
08: stressberry-plot resFile.dat -o both.png --not-transparent -f
```

Fig. 2. Installation and Execution of Stressberry.

Fig. 3 depicts the results that we obtained by running the Stressberry test on the RPi Zero W, RPi Zero 2 W, RPi 3B, RPi 3B+, and RPi 4B. The curves can be divided into three parts: (1) a 150-second idle portion for phase 2, (2) a 300-second stress activity for phase 3, and (3) a 150-second idle portion for phase 4. When running Stressberry without specifying the number of cores, the tools will activate all of them. The RPi Zero W just has one core, and its maximum temperature went barely over 45°C during the stress activity. All the other RPis have four cores, that were activated by Stressberry in this experiment. According to this test, the RPi 3B+ has the highest baseline temperature, while the RPi 3B has the highest temperature during the stress period. It is noted that all the tests of Fig. 3 correspond to naked RPis (no cases), and without any cooling solutions. That is, the small fan was not used for the RPi 4B in this experiment.

The goal of Fig. 4 is to show how a cooling solution can significantly improve the temperature of the CPU. This experiment was conducted with an RPi 4B. The first curve (in orange) corresponds to running Stressberry without any additional heat control system, while a small fan was used for the second curve (in black). The minimum difference is greater than 10°C, and in the best case, it is almost 25°C. It is worth mentioning that the results will significantly vary with the chosen cooling solution (heat sink, fan, a combination of heat sink and fan, specific cases, etc.) and its size.

Fig. 5 aims to determine the impact on the temperature of an RPi, when overclocked or not. The experiment was conducted by running Stressberry on an RPi Zero 2 W. The figure has four curves, two for the RPi not overclocked (temperature and CPU frequency) and two for the RPi overclocked at 1300 MHz (temperature and CPU frequency). In the initial 150-second idle portion, the temperature was the same since the CPU frequency mainly stayed at 600 MHz for both cases. In the 300-second portion of stress, the CPU frequency was steady at 1000 MHz for the RPi not overclocked. However, when overclocked, the CPU frequency started at 1300 MHz and went down to 1000 MHz in the intent to control the rising temperatures. In the final 150-second idle portion, the CPU frequency mainly stayed at 600 MHz for both cases, allowing the temperatures to fall down toward the baseline temperature.

Fig. 6 shows the effect on the temperature of a RPi, when varying the number of cores. The experiment was carried out by running Stressberry on an RPi Zero 2 W not overclocked, activating 1, 2, 3, and 4 cores, respectively. As expected, when under stress, the temperatures get higher with the increasing number of cores.

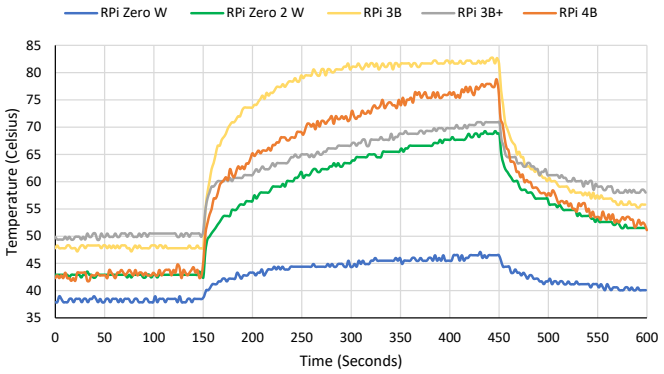


Fig. 3. Results of Stressberry for Different RPi Models.

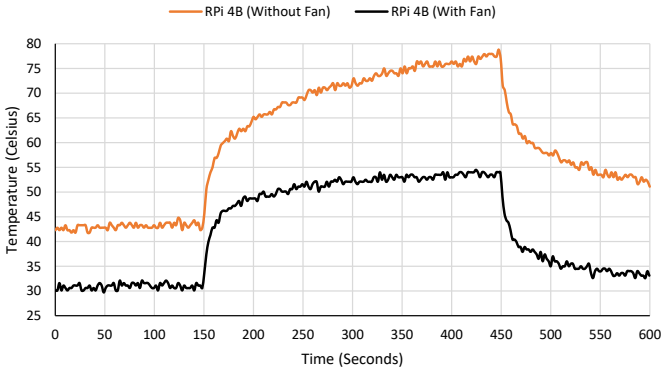


Fig. 4. Results of Stressberry for an RPi 4B with/without Fan.

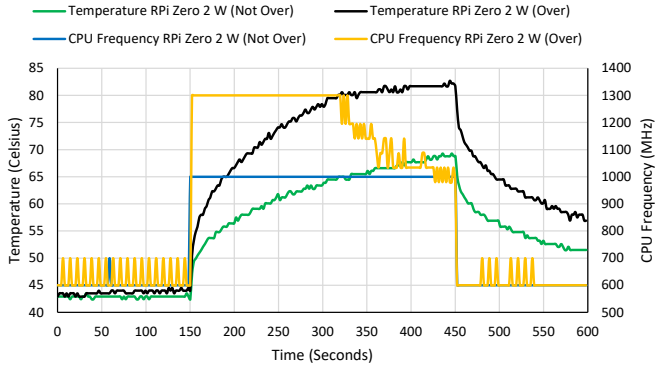


Fig. 5. Results of Stressberry for the RPi Zero 2 W when Overclocked and when Not.

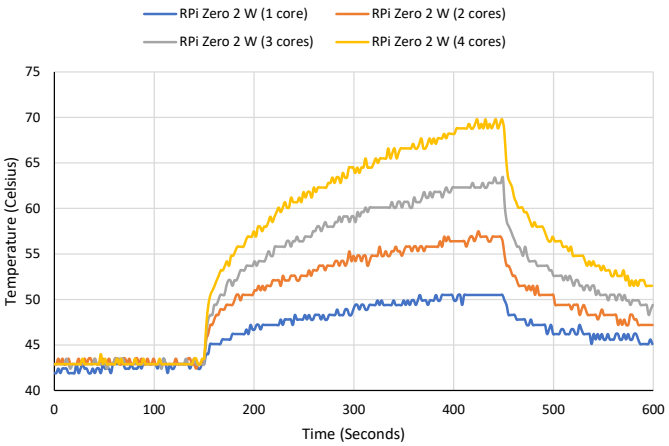


Fig. 6. Results of Stressberry for the RPi Zero 2 W when Varying the Number of Cores.

B. CPU Performance with 7-Zip

The 7-Zip archiving tool [26][27] can be used to pack and compress files into archives, as well as to extract the files from archive formats such as ZIP or 7z. It also has a benchmarking tool built inside it, that assesses the power of a CPU through the LZMA [28] (Lempel–Ziv–Markov chain Algorithm) compression and decompression. The tool reports how fast a CPU processes the compression and decompression instructions over dummy data, displaying the results in MIPS (Million Instructions Per Second). Fig. 7 shows the instructions that we used to install and execute the 7-Zip benchmark. Line 01 installs the tool, while the manual is consulted in Line 02. Lines 03-05 run the benchmark for 1, 2, and 4 threads, respectively. Since the RPi Zero W only has one core, Lines 04-05 were not executed on this RPi.

```
01: apt-get install p7zip
02: man 7zr
03: 7zr b -mmt1
04: 7zr b -mmt2
05: 7zr b -mmt4
```

Fig. 7. Installation and Execution of the Benchmark of 7-Zip.

Fig. 8 and 9 depict the results obtained for the CPU assessment with 7-Zip for compression and decompression, respectively. The dictionary size was 2^{23} = 8 MB. There are two results for the RPi Zero 2 W: one without overclocking, and the other with overclocking the device at 1.3 GHz. It is significantly noticeable how the RPi 4B outperformed all the other RPis. Overclocking the RPi Zero 2 W boots its CPU performance. The RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ have similar performance.

C. CPU Performance with Sysbench

Sysbench [29] is a scriptable multi-threaded benchmark tool based on LuaJIT. Sysbench has several tests (`cpu`, `memory`, `fileio`, `threads`, and `mutex`) for the CPU performance, memory speed, file I/O access, threads subsystem performance, and mutex performance, respectively.

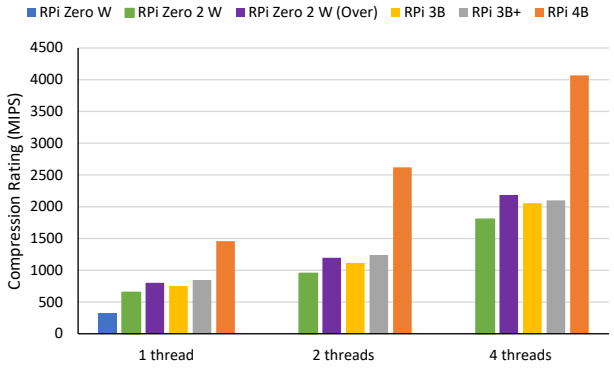


Fig. 8. Compression Rating with 7-Zip for Different RPi Models.

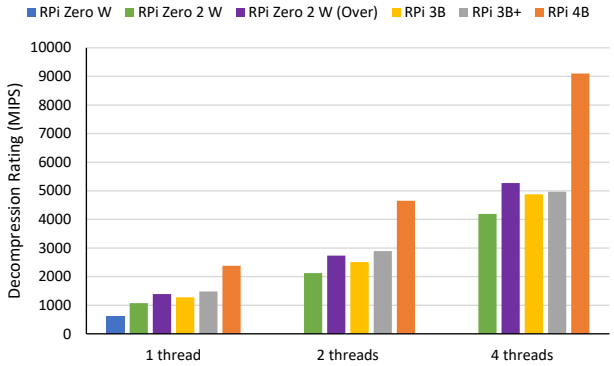


Fig. 9. Decompression Rating with 7-Zip for Different RPi Models.

In this experiment, we evaluated the CPU performance of the different RPis with Sysbench. The test consists in generating random numbers and verifying if they are prime or not, by doing standard divisions of any selected number by all integers between 2 and the square root of this number. If any division gives a remainder of 0, Sysbench starts over by generating a new random number and trying again. The results are reported in events/sec. Fig. 10 gives the instructions that we used to install are run the CPU performance test. Sysbench was installed with Line 01, and the manual was consulted in Lines 02-03. Lines 04-06 run the CPU performance test for 1, 2, and 4 threads, respectively. We limited the total execution time to 20 seconds, and the randomly generated numbers to be tested were inferior to 20,000, as specified by the options of the commands.

```
01: apt-get install sysbench
02: sysbench --help
03: sysbench cpu help
04: sysbench cpu --threads=1 --time=20 --cpu-max-prime=20000 run
05: sysbench cpu --threads=2 --time=20 --cpu-max-prime=20000 run
06: sysbench cpu --threads=4 --time=20 --cpu-max-prime=20000 run
```

Fig. 10. Installation and Execution of Sysbench to Assess the CPU Performance.

Fig. 11 depicts the results that we obtained for this test. It is noted that the RPi Zero W has a result only for one thread. This test confirms the results obtained in Section IV.B with 7-Zip, where the RPi 4B dramatically outperformed the other RPis. Overclocking the RPi Zero 2 W improved its performance. Moreover, the RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ have a similar performance.

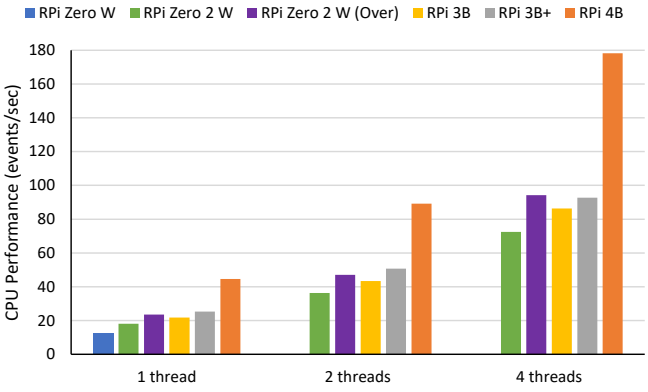


Fig. 11. CPU Performance with Sysbench for Different RPi Models.

D. Memory Speed with Geekbench and STREAM

We initially utilized Sysbench [29] to benchmark the read and write access performance of the memory (RAM) of the RPis. However, we did not get consistent results. Similar problems were reported in [30]. Hence, we opted to use Geekbench [31] and STREAM [32][33].

Geekbench [31] is a cross-platform benchmark program (Windows, macOS, Linux i386/amd64, Android, iOS, etc.) that reports performance related to the integer arithmetic, floating-point arithmetic, and memory. It is a commercial product from Primate Labs; however, a limited version can be downloaded and used for free. At the level of the memory assessment, the older versions of Geekbench [34] has four convenient tests: (1) "Read Sequential" that loads values from memory into registers, (2) "Write Sequential" that stores values from registers into memory, (3) "Stdlib Write" that writes a constant value to a block of memory using functions from the C Standard Library (`memset`), and (4) "Stdlib Copy" that copies values from one block of memory to another using functions from the C Standard Library (`memcpy`). Fig. 12 shows the results that we obtained for "Read Sequential", "Write Sequential", "Stdlib Write", and "Stdlib Copy", respectively, in MB/sec. The RPi Zero W has the poorest performance. The experiment seems to indicate that the RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ have a similar memory access, for both reading and writing. The RPi 4B significantly outperformed the other devices under test.

Fig. 12. Results of the Memory Performance with Geekbench for Different RPi Models.

STREAM [32][33] is a synthetic benchmark designed to measure sustainable memory bandwidth and the corresponding computation rate for four simple vector kernels: Copy, Scale, Add, and Triad. Copy just transfers data from one memory location to another, i.e., copies it (A[i] = B[i]). Scale takes the value of the first location and multiplies it with a certain constant, before storing it in a second place, i.e., scales it (A[i] = m*B[i]). Add reads data from two different locations, adds them up and writes the result to a third place (A[i] = B[i] + C[i]). Triad reads data from a first location, scales it, then adds data from a second one and writes to a third place (A[i] = m*B[i] + C[i]). Fig. 13 displays the steps that we followed to download, compile, and run STREAM. Notice that STREAM can be compiled with or without OpenMP [35]–[37] support. Lines 01-02 installed Git (a revision control system) and cloned the STREAM repository, respectively. In Line 04, we compiled `stream.c` and specified that the elements of the unidimensional arrays are floating-point numbers in double precisions (`double`), the size of the arrays is 10 MB, each kernel should be run ten times, and activated support for OpenMP. Line 05 is optional. If not specified, one thread is activated for the RPi Zero W, and four threads for the other devices under test.

```
01: apt-get install git
02: git clone https://github.com/jeffhammond/STREAM.git
03: cd STREAM
04: gcc -o stream-bin -O3 -march=native -DSTREAM_TYPE=double \
 -DSTREAM_ARRAY_SIZE=10000000 -DNTIMES=10 -fopenmp stream.c
05: export OMP_NUM_THREADS=<NUM_CPU_CORES>
06: ./stream-bin
```

Fig. 13. Download, Compilation, and Execution of STREAM to Assess the Memory Performance.

Fig. 14 depicts the assessment of the different RPis with STREAM, using one thread for the RPi Zero W and four threads for the other RPis. It is noticeable how the RPi 4B outperformed the other SBCs. The RPi Zero W is much slower. The other devices under test performed similarly.

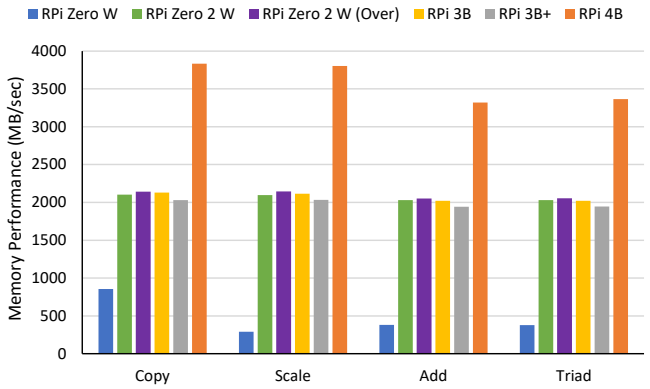


Fig. 14. Results of the Memory Performance with STREAM for Different RPi Models.

E. Sequential Access Performance on the MicroSD Card with Sysbench

The goal of this experiment was to evaluate the sequential read and write access performance on the microSD card for the different RPis with Sysbench [29], for one thread. Fig. 15 shows the instructions that we used to run the file access

performance test, for sequential readings. The manual was consulted in Lines 01-02. Line 03 generated 32 files of size 64 MiB each, for a total of 2 GiB. Those files are the ones from which Sysbench performed the read operations during the test. In Line 04, the test was run for a block size of 1 kiB, for 20 seconds. This line was executed several times, where the block size was varied to 2, 8, 32, 64, 128, 256, 512, 1024, and 2048 kiB, respectively. Finally, Line 05 erased all the 32 temporary files created in Line 03.

```
01: sysbench --help
02: sysbench fileio help
03: sysbench fileio --file-num=32 --file-total-size=2G prepare
04: sysbench fileio --file-num=32 --file-total-size=2G \
 --file-extra-flags=direct --file-test-mode=seqrd \
 --file-block-size=1k --time=20 run
05: sysbench fileio --file-num=32 --file-total-size=2G cleanup
```

Fig. 15. Execution of Sysbench to Evaluate the Sequential Read Assess Performance on the microSD Card.

Fig. 16 presents the instructions used to benchmark the sequential write access performance on the microSD card, for the different RPis with Sysbench, for one thread. In this case, there is no need to pre-generate temporary files since Sysbench will not read, but write. In Line 01 the test was run for a block size of 1 kiB, for 20 seconds, writing files (up to 32 files with a maximum size of 64 MiB each). This line was executed several times, where the block size was varied to 2, 8, 32, 64, 128, 256, 512, 1024, and 2048 kiB, respectively. Finally, Line 02 removed all the files created in Line 01.

```
01: sysbench fileio --file-num=32 --file-total-size=2G \
 --file-extra-flags=direct --file-test-mode=seqwr \
 --file-block-size=1k --time=20 run
02: sysbench fileio --file-num=32 --file-total-size=2G cleanup
```

Fig. 16. Execution of Sysbench to Evaluate the Sequential Write Assess Performance on the microSD Card.

Fig. 17 and 18 depict the results that we obtained for the sequential read and write access performance, respectively, for the different RPis. The RPi Zero W has the poorest performance, while the RPi 4B has the best one. The RPi Zero 2 W (overclocked or not), RPi 3B, and RPi 3B+ have a comparable performance.

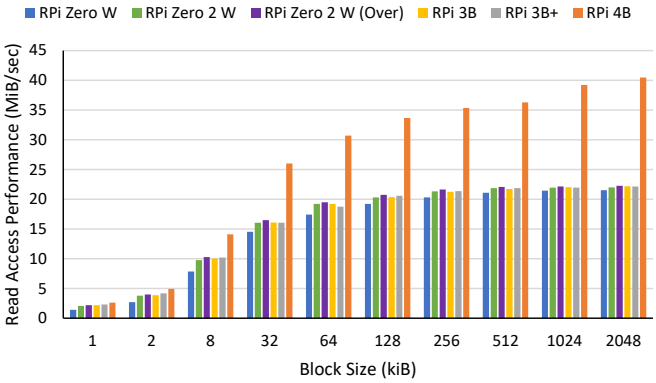


Fig. 17. Sequential Read Access Performance on the microSD Card with Sysbench for Different RPi Models.

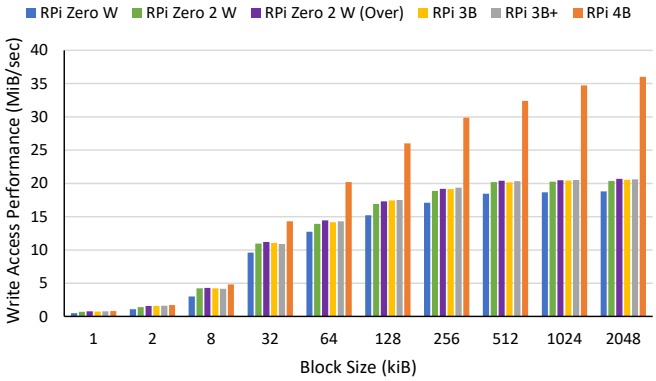


Fig. 18. Sequential Write Access Performance on the microSD Card with Sysbench for Different RPi Models.

F. Sequential Write Access Performance on the microSD Card with dd

This experiment aimed to evaluate the sequential write access performance on the microSD card for the different RPis with "dd", for one thread. The command "dd" is a standard Unix/Linux tool to convert and copy files. Fig. 19 shows the instructions that we used to run the sequential write access performance test with "dd". The manual was consulted in Line 01. In Line 02, the test was run. Here, the tool created a 512 kiB file by making 512 write operations, each one with a block size of 1 kiB. This last line was executed several times, where the block size was changed to 2, 8, 32, 64, 128, 256, 512, 1024, and 2048 kiB, respectively, creating files of 1, 4, 16, 32, 64, 128, 256, 512, and 1024 MiB, respectively.

```
01: man dd
02: dd if=/dev/zero of=/home/pi/test bs=1k count=512 oflag=direct
```

Fig. 19. Execution of "dd" to Evaluate the Sequential Write Assess Performance on the microSD Card.

Fig. 20 depicts the results that we obtained for the different RPis. It is very similar to the performance results of Fig. 18.

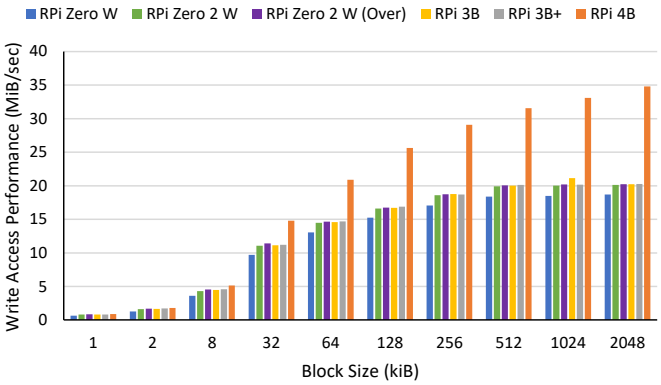


Fig. 20. Sequential Write Access Performance on the microSD Card with "dd" for Different RPi Models.

G. Audio Conversion Performance with Phoronix Test Suite

Phoronix Test Suite [38] (PTS) is a free and open-source framework for conducting automated performance tests. It is multiplatform, and the default version has more than 600 individual test profiles and more than 200 test suites, covering

a wide range of applications such as audio format conversions, encryption and decryption algorithms, compression and decompression algorithms, timed compilation of widely-used open-source software, memory access performance, file access performance, etc. The framework is designed to be extensible so that new test profiles and suites can be easily added.

We used PTS [38] to evaluate the performance of the RPis when converting/encoding sample WAV files to MP3 (encode-mp3), sample WAV files to FLAC (encode-flac), sample WAV files to Monkey's Audio APE (encode-ape), sample WAV files to WavPack (encode-wavpack), and sample WAV files to Opus (encode-opus). Fig. 21 shows the instructions that we used to install and run the audio conversion performance tests with PTS. First, the dependencies were installed in Line 01. In Line 02, PTS was cloned from GitHub. Lines 04-08 run the different conversion tests.

```
01: apt-get install php-cli php-xml git
02: git clone https://github.com/phoronix-test-suite/\
 phoronix-test-suite.git
03: cd phoronix-test-suite
04: ./phoronix-test-suite benchmark encode-mp3
05: ./phoronix-test-suite benchmark encode-flac
06: ./phoronix-test-suite benchmark encode-ape
07: ./phoronix-test-suite benchmark encode-wavpack
08: ./phoronix-test-suite benchmark encode-opus
```

Fig. 21. Installation and Execution of Phoronix Test Suite to Evaluate the Performance of Audio Conversions.

Fig. 22 depicts the audio conversion performance results in seconds. The lower is the conversion time, the better/faster is the RPi. The RPi Zero W had the longest conversion time, while the RPi 4B had the shortest. The RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ had similar performance.

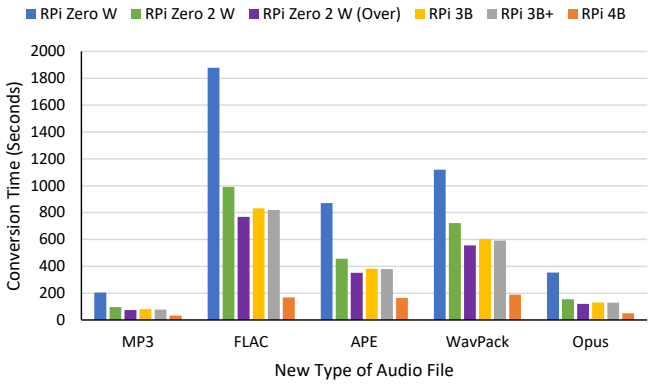


Fig. 22. Audio Conversion Performance with PTS for Different RPi Models.

H. Other Performance Evaluation with Phoronix Test Suite

We also utilized PTS [38] for other performance evaluations. The "`openssl`" test of PTS assesses (1) the number of digital signatures that can be performed per second and (2) the number of verifications of digital signatures that can be processed per second, using RSA with 4096-bit keys. Table I has the results that we obtained for the different RPis. In this test also, the RPi 4B had a much better performance. The RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ had comparable performances.

TABLE I. SIGNING AND VERIFYING PERFORMANCE WITH PTS

Test	Signing (sign/sec)	Verifying (verify/sec)
RPi Zero W	2.6	176.1
RPi Zero 2 W	39.4	2834.7
RPi Zero 2 W (Over)	47.2	3380.2
RPi 3B	45.9	3329.3
RPi 3B+	46.4	3337.5
RPi 4B	120.0	9126.4

The total time required to compile an open-source software is considered a good benchmark to measure the performance of computers. PTS [38] offers several profiles to get a timed compilation of well-accepted software. Table II shows the results that we obtained to compile ImageMagick [39][40] (an application to create, edit, compose, or convert digital images) and MPlayer [41] (a movie player), using PTS. The big difference between the RPi Zero W and the other RPis can be partially explained by the options used with the "`make`" utility. PTS used "`make -j1`" for the RPi Zero W, and "`make -j4`" for the other RPis. This option specifies the number of jobs (commands) that can be run simultaneously during the compilation process.

TABLE II. TOTAL TIME REQUIRED TO COMPILE OPEN-SOURCE SOFTWARE WITH PTS

Test	ImageMagick (Time in Seconds)	MPlayer (Time in Seconds)
RPi Zero W	7350.6	12990.6
RPi Zero 2 W	825.7	1390.5
RPi Zero 2 W (Over)	758.2	1196.0
RPi 3B	773.9	1241.1
RPi 3B+	762.2	1204.4
RPi 4B	388.4	507.6

I. TCP Throughput with Iperf

The goal of this experiment is to determine the maximum TCP throughput that can be obtained between two end-devices connected as specified in Fig. 1, where at least one RPi is utilized as an end-device. To do so, we used Iperf [42][43], a free, open-source command-line tool for network performance measurement between two network devices. It is based on the client/server model, and reports parameters such as the throughput, delay jitter, and packet loss. Iperf v2.0.14a is available as a pre-compiled package from the Raspberry Pi OS repositories. However, this version has limitations. Hence, we downloaded and installed a newer version (Iperf v2.1.n) for our experiments. We used Iperf to determine the "maximum" TCP throughput between the client and the server. When using this test, the client tries to overwhelm the server by creating a unidirectional TCP flow (from the client to the server) and sending as many segments as allowed by the TCP control mechanism (congestion window). At the end of the experiment, by default 10 seconds, the TCP throughput is displayed.

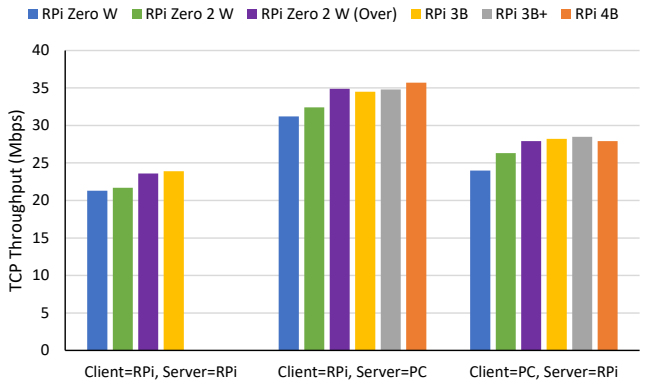


Fig. 23. TCP Throughput with Iperf when Varying the Roles of the RPis for Different RPi Models.

Fig. 23 depicts the TCP throughput that we obtained in three different scenarios: (1) both the client and server were RPis of the same models, (2) the client was an RPi while the server was a PC, and (3) the client was a PC while the server was an RPi. The specifications of the PC were given in Section III.C. It is worth noting that in the first group of bars, there are only four bars, corresponding to two RPi Zero W (blue bar), two RPi Zero 2 W not overclocked (green bar), two RPi Zero 2 W overclocked (purple bar), and two RPi 3B (yellow bar). There are no bars for the RPi 3B+ and the RPi 4B, since we only had one of each of these SBCs. For this experiment, we set up the wireless router to use a maximum bandwidth of 145 Mbps in the 2.4 GHz band. It is noted that all the RPis had a bitrate that capped out at 72.2 Mbps in this band.

The best performance is obtained when the server is a PC (second group of bars), while the worst corresponds to the case of using two RPis of the same models (first group of bars) for the client and server. Since the PC that we used is more potent than the RPis, when used as a server, it is much faster to discard the received segments from the client, and then to reopen its TCP congestion window, allowing the client to send the TCP segments at a higher rate, resulting in a better performance for this case.

The RPi 3B+ and RPi 4B are dual-bands. Hence, we also made some performance evaluations of the TCP throughput in the 5 GHz band, by setting up the wireless router to use a maximum bandwidth of 867 Mbps in the 5 GHz band. It is worth mentioning that the two RPis (RPi 3B+ and RPi 4B) had a bitrate that capped out at 433.3 Mbps in this band. Table III has the results that we obtained. By changing from the 2.4 GHz to the 5 GHz band, the improvement in throughput is noticeable.

TABLE III. TCP THROUGHPUT WITH IPERF IN DIFFERENT BANDS FOR THE RPI 3B+ AND RPI 4B

Test	2.4 GHz	5 GHz
Client=RPi 3B+, Server=PC	34.80 Mbps	102.10 Mbps
Client=PC, Server=RPi 3B+	28.50 Mbps	81.40 Mbps
Client=RPi 4B, Server=PC	35.70 Mbps	104.25 Mbps
Client=PC, Server=RPi 4B	27.90 Mbps	85.70 Mbps

J. TCP Latency with our Own Benchmark

The aim of this experiment is to determine the TCP latency to send a specific TCP payload from a source to a destination. To accomplish this objective, we wrote our own benchmark and used the testbed of Fig. 1, with an RPi as the source and a PC as a destination. The specifications of the PC were given in Section III.C. For this experiment, we set up the wireless router to use a maximum bandwidth of 145 Mbps in the 2.4 GHz band. Fig. 24 shows the results that we got for the TCP latency when varying the payload to be transmitted through TCP from 100 to 10,000 bytes. The RPi Zero W has the longest latency, followed by the RPi Zero 2 W not overclocked. The RPi Zero 2 W overclocked, RPi 3B, RPi 3B+, and RPi 4B had a similar performance.

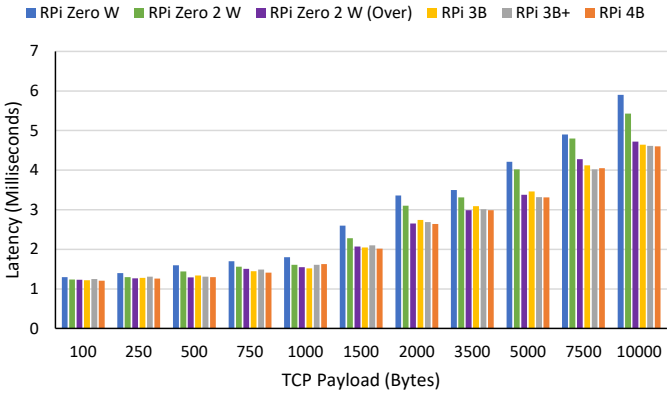


Fig. 24. TCP Latency with our Own Benchmark for Different RPi Models.

V. CONCLUSION AND FUTURE WORK

In this paper, we evaluated the performance of a number of SBCs: RPi Zero W, RPi Zero 2 W, RPi Zero 2 W overclocked, RPi 3B, RPi 3B+, and RPi 4B. The RPi 4B significantly outperformed the other SBCs under test. So far, it is the most potent RPis released by the Raspberry Pi Foundation, with a basic price (board only) of US$35, US$45, US$55, and US$75 for the 1, 2, 4, and 8 GB models, respectively. If used at its maximum power, a cooling solution is recommended; otherwise, the CPU will be throttled, resulting in slowing down the CPU frequency. Our study showed that the RPi Zero, released in February 2017, has limited capacity. It is the only 32-bit processor of the study, with just one core. However, it might still be the solution for many projects with low CPU and RAM requirements. Its basic price of US$10 (board only) is very attractive for projects with a low budget. In general, the RPi Zero 2 W overclocked, RPi 3B, and RPi 3B+ had similar performances. In some tests, the RPi Zero 2 W overclocked had a light advantage, while the RPi Zero 3B+ was slightly better in other tests. The basic price (board only) for the RPi Zero 2 W, RPi 3B, and RPi 3B+ is US$15, US$35, and US$35, respectively. The RPi Zero 2 W is the last SBC released by the Raspberry Pi Foundation, with this amazing price. It has the same form factor as the RPi Zero W, allowing an easy upgrade when required. However, its limited RAM (512 MB) can be a limitation for some projects, compared to the 1 GB of the RPi 3B and RPi 3B+.

When Ethernet is a project requirement, the RPi 3B, RPi 3B+, and RPi 4B have an integrated port. They offer Fast Ethernet (100 Mbps), Gigabit Ethernet over USB 2.0 (maximum throughput 300 Mbps), and Gigabit Ethernet, respectively. For the other SBCs of this study, Ethernet can still be added through a USB port. However, this solution should be considered only in existing projects as an extension, since new projects will be more cost-effective when using RPis with native Ethernet support.

The RPi 400 [44] was not studied in the paper. It is a keyboard that incorporates an RPi 4B into it, with minor modifications. It is the easiest way to build a desktop computer based on an RPi. Even if they are mostly identical, the Raspberry Pi Foundation releases the RPi 400 with its Broadcom BCM2711 processor clocked at 1.8 GHz, while the RPi 4B was set to 1.5 GHz in previous versions of the Raspberry Pi OS. This is just due to an integrated robust cooling solution in the keyboard of the RPi 400. However, the last version of the Raspberry Pi OS (October 2021) now also sets the Broadcom BCM2711 processor of the RPi 4B at 1.8 GHz. Hence, as specified in Section III.A, the CPU of our RPi 4B was set to 1.8 GHz for all our experiments.

The RPi 3B+ is an improved version of the RPi 3B. The main difference is in the upgraded support for the network connection. According to the Raspberry Pi Foundation website, the RPi 3B will remain in production until at least January 2026 [22]. Both (RPi 3B and RPi 3B+) have the same price: US$35 for the board only. In most of our experiments, they had a similar performance with a slight advantage for the RPi 3B+. Hence, for new projects that are planning to use the RPi 3B, our recommendation is to consider the RPi 3B+ instead.

Even though a powerful processor means more heat and higher temperatures when running intensive CPU workloads, an interesting finding of this study is that the RPi 3B reached the highest temperature (more than 80º C) during the stress activity, which is significantly higher than the temperature reached by the RPi 3B+ and the RPi 4B for the same CPU workload. This shows that these two later models might have been improved in terms of heat management in comparison to their predecessor (RPi 3B).

As future work, we plan to extend our study to other SBCs, such as the ones of BeagleBoard [45]. We are also interested in focusing our efforts on the network performance of IPv4 and IPv6, for both Ethernet and WiFi, in SBCs.

ACKNOWLEDGMENT

We are grateful to “Faculty Commons” and the “College of Science & Mathematics” at Jacksonville State University for partially funding this project.

REFERENCES

[1] W. Jolles, “Broad-scale Applications of the Raspberry Pi: A Review and Guide for Biologists,” Methods Ecol. Evol., vol. 12, no. 9, pp. 1562–1579, 2021, doi: 10.1111/2041-210X.13652.

[2] L. K. Ramasamy and S. Kadry, Blockchain in the Industrial Internet of Things. IOP Publishing Ltd, 2121.

[3] D. B. C. Lima, R. M. B. Da Silva Lima, D. De Farias Medeiros, R. I. S. Pereira, C. P. De Souza, and O. Baiocchi, “A Performance Evaluation of Raspberry Pi Zero W Based Gateway Running MQTT Broker for IoT,” in Proceedings of the 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON 2019), Oct. 2019, pp. 0076–0081, doi: 10.1109/IEMCON.2019.8936206.

[4] A. W. Daher, A. Rizik, M. Muselli, H. Chible, and D. D. Caviglia, “Porting Rulex Software to the Raspberry Pi for Machine Learning Applications on the Edge,” Sensors, vol. 21, no. 19, pp. 1–16, 2021, doi: 10.3390/s21196526.

[5] A. Komninos, I. Simou, N. Gkorgkolis, and J. Garofalakis, “Performance of Raspberry Pi Microclusters for Edge Machine Learning in Tourism,” in Proceedings of the 2019 European Conference on Ambient Intelligence (AmI 2019), Nov. 2019, vol. 2492, pp. 1–10.

[6] H. D. Ghael, L. Solanki, and G. Sahu, “A Review Paper on Raspberry Pi and its Applications,” Int. J. Adv. Eng. Manag., vol. 2, no. 12, pp. 225–227, 2021.

[7] R. Morabito, “A Performance Evaluation of Container Technologies on Internet of Things Devices,” in Proceedings of the 2016 IEEE Conference on Computer Communications Workshops (INFOCOM 2016), Apr. 2016, pp. 1–2, doi: 10.1109/INFCOMW.2016.7562228.

[8] A. K. Kyaw, Y. Chen, and J. Joseph, “Pi-IDS: Evaluation of Open-source Intrusion Detection Systems on Raspberry Pi 2,” in Proceedings of the 2015 2nd International Conference on Information Security and Cyber Forensics (InfoSec 2015), Nov. 2015, pp. 165–170, doi: 10.1109/InfoSec.2015.7435523.

[9] A. Aspernäs and T. Simonsson, “IDS on Raspberry Pi : A Performance Evaluation,” 2015. http://lnu.diva-portal.org/smash/get/diva2:819555/FULLTEXT01.pdf.

[10] E. Gamess and S. Hernandez, “Performance Evaluation of SNMPv1/2c/3 using Different Security Models on Raspberry Pi,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 1–9, 2021, doi: 10.14569/ijacsa.2021.0121101.

[11] D. Mauro and K. Schmidt, Essential SNMP, 2nd ed. O’Reilly Media, 2005.

[12] L. H. Nunes et al., “Performance and Energy Evaluation of RESTful Web Services in Raspberry Pi,” in Proceedings of the 2014 IEEE 33rd International Performance Computing and Communications Conference (IPCCC 2014), Dec. 2014, pp. 1–9, doi: 10.1109/PCCC.2014.7017086.

[13] Y. Guamán, G. Ninahualpa, G. Salazar, and T. Guarda, “Comparative Performance Analysis between MQTT and CoAP Protocols for IoT with Raspberry Pi 3 in IEEE 802.11 Environments,” in Proceedings of the 2020 15th Iberian Conference on Information Systems and Technologies (CISTI 2020), Jun. 2020, pp. 1–6, doi: 10.23919/CISTI49556.2020.9140905.

[14] A. Banks and R. Gupta, “MQTT Version 3.1.1,” OASIS Standard, Oct. 2014. http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.pdf.

[15] “MQTT Homepage.” https://mqtt.org.

[16] “CoAP – Constrained Application Protocol Homepage.” https://coap.technology.

[17] Z. Shelby, K. Hartke, and C. Bormann, “The Constrained Application Protocol (CoAP),” RFC 7252. Internet Engineering Task Force (IETF), Jun. 2014.

[18] D. Hawthorne, M. Kapralos, R. W. Blaine, and S. J. Matthews, “Evaluating Cryptographic Performance of Raspberry Pi Clusters,” in Proceedings of the 2020 IEEE High Performance Extreme Computing Conference (HPEC 2020), Sep. 2020, pp. 1–9, doi: 10.1109/HPEC43674.2020.9286247.

[19] E. Fernando, D. Agustin, M. Irsan, D. F. Murad, H. Rohayani, and D. Sujana, “Performance Comparison of Symmetries Encryption Algorithm AES and DES with Raspberry Pi,” in Proceedings of the 2019 4th International Conference on Sustainable Information Engineering and Technology (SIET 2019), Sep. 2019, pp. 353–357, doi: 10.1109/SIET48054.2019.8986122.

[20] “Raspberry Pi Zero W – Raspberry Pi.” https://www.raspberrypi.com/products/raspberry-pi-zero-w.

[21] “Raspberry Pi Zero 2 W – Raspberry Pi.” https://www.raspberrypi.com/products/raspberry-pi-zero-2-w.

[22] “Raspberry Pi 3 Model B – Raspberry Pi.” https://www.raspberrypi.com/products/raspberry-pi-3-model-b.

[23] “Raspberry Pi 3 Model B+ – Raspberry Pi.” https://www.raspberrypi.com/products/raspberry-pi-3-model-b-plus.

[24] "Raspberry Pi 4 Model B – Raspberry Pi." https://www.raspberrypi.com/products/raspberry-pi-4-model-b.

[25] "Stressberry: Stress Tests for the Raspberry Pi." https://github.com/nschloe/stressberry.

[26] "7-Zip." https://www.7-zip.org.

[27] D. S. N. Nunes, F. A. Louza, S. Gog, M. Ayala-Rincon, and G. Navarro, "A Grammar Compression Algorithm based on Induced Suffix Sorting," in Proceedings of the 2018 Data Compression Conference, Mar. 2018, pp. 42–51, doi: 10.1109/DCC.2018.00012.

[28] A. R. Rosete, K. R. Baker, and Y. Ma, "Using LZMA Compression for Spectrum Sensing with SDR Samples," in Proceedings of the 2018 9th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON 2018), Nov. 2018, pp. 282–287, doi: 10.1109/UEMCON.2018.8796574.

[29] "Sysbench: Scriptable Database and System Performance Benchmark." https://github.com/akopytov/sysbench.

[30] "Raspberry Pi 4 B Review and Benchmark - What's Improved over Pi 3 B+," Sep. 2019. https://ibug.io/blog/2019/09/raspberry-pi-4-review-benchmark.

[31] "Geekbench 5 - Cross-Platform Benchmark." https://www.geekbench.com.

[32] "STREAM: Sustainable Memory Bandwidth in High Performance Computers." https://www.cs.virginia.edu/stream.

[33] "STREAM: The de Facto Industry Standard Benchmark for Measuring Sustained Memory Bandwidth." https://github.com/jeffhammond/STREAM.

[34] "GeekBench for Linux ARM - Primate Labs Support." http://support.primatelabs.com/discussions/geekbench/70-geekbench-for-linux-arm.

[35] B. R. de Supinski et al., "The Ongoing Evolution of OpenMP," Proc. IEEE, vol. 106, no. 11, pp. 2004–2019, 2018, doi: 10.1109/JPROC.2018.2853600.

[36] OpenMP Architecture Review Board, OpenMP Application Programming Interface Specification Version 5.1. Independently published, 2020.

[37] T. Katagiri, "Basics of OpenMP Programming," in The Art of High Performance Computing for Computational Science, Vol. 1, M. Geshi, Ed. Springer, 2019, pp. 45–59.

[38] "Phoronix Test Suite: Open-Source, Automated Benchmarking." https://www.phoronix-test-suite.com.

[39] M. Still, The Definitive Guide to ImageMagick, 1st ed. Apress, 2006.

[40] S. ML, A. PJ, and S. DN, "Document Image Analysis Using Imagemagick and Tesseract-ocr," IARJSET, vol. 3, no. 5, pp. 108–112, 2016, doi: 10.17148/iarjset.2016.3523.

[41] "MPlayer: The Movie Player." http://www.mplayerhq.hu/design7/info.html.

[42] "IPerf2: A Tool that Measures Network Performance of TCP/UDP Including Latency." https://sourceforge.net/projects/iperf2.

[43] V. J. D. Barayuga and W. E. S. Yu, "Packet Level TCP Performance of NAT44, NAT64 and IPv6 using Iperf in the Context of IPv6 Migration," in Proceedings of the 2015 5th International Conference on IT Convergence and Security (ICITCS 2015), Aug. 2015, pp. 1–3, doi: 10.1109/ICITCS.2015.7293006.

[44] "Raspberry Pi 400 Personal Computer Kit – Raspberry Pi." https://www.raspberrypi.com/products/raspberry-pi-400.

[45] "BeagleBoard.org - Community Supported Open Hardware Computers for Making." https://beagleboard.org.